A GUIDE to the development implementation and evaluation of clinical practice guidelines.

How to use the evidence: assessment and application of scientific evidence

NHMRC

National Health and Medical Research Council

How to use the evidence: assessment and application of scientific evidence

Handbook series on preparing clinical practice guidelines

Endorsed February 2000

NHMRC

National Health and Medical Research Council

ISBN 0 642 43295 3

The strategic intent of the National Health and Medical Research Council (NHMRC) is to work with others for the health of all Australians, by promoting informed debate on ethics and policy, providing knowledge-based advice, fostering a high quality and internationally recognised research base, and applying research rigour to health issues.

NHMRC documents are prepared by panels of experts drawn from appropriate Australian academic, professional, community and government organisations. NHMRC is grateful to these people for the excellent work they do on its behalf. The work is usually performed on an honorary basis and in addition to their usual work commitments.

This document is sold through Government Info Bookshops at a price that covers the cost of printing and distribution only. For publication purchases please contact AusInfo on their toll-free number 132 447, or through their Internet address:
www.ausinfo.gov.au/general/gen_hottobuy.htm

Production by Biotext, Canberra

CONTENTS

TABLES

FIGURES

BOXES

PREFACE

Clinical practice guidelines are systematically developed statements that assist clinicians, consumers and policy makers to make appropriate health care decisions. Such guidelines present statements of ‘best practice’ based on a thorough evaluation of the evidence from published research studies on the outcomes of treatment or other health care procedures. The methods used for collecting and evaluating evidence and developing guidelines can be applied to a wide range of clinical interventions and disciplines, including the use of technology and pharmaceuticals, surgical procedures, screening procedures, lifestyle advice, and so on.

In 1995, recognising the need for a clear and widely accessible guide for groups wishing to develop clinical practice guidelines, the National Health and Medical Research Council (NHMRC) published a booklet to assist groups to develop and implement clinical practice guidelines. In 1999 a revised version of this booklet was published called *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999), which includes an outline of the latest methods for evaluating evidence and developing and disseminating guidelines.

The emerging guideline processes are complex, however, and depend on the integration of a number of activities, from collection and processing of scientific literature to evaluation of the evidence, development of evidence-based recommendations or guidelines, and implementation and dissemination of the guidelines to relevant professionals and consumers. The NHMRC has therefore decided to supplement the information in the guideline development booklet (NHMRC 1999) with a series of handbooks with further information on each of the main stages involved. Experts in each area were contracted to draft the handbooks. An Assessment Panel was convened in June 1999 to oversee production of the series. Membership of the Assessment Panel and the writing group for this handbook are shown at Appendix A.

Each of the handbooks in the series focuses on a different aspect of the guideline development process (review of the literature, evaluation of evidence, dissemination and implementation, consumer publications, economic assessment and so on). This handbook focuses on how to evaluate and use the evidence gathered from a systematic literature review to inform the development of evidence-based clinical guidelines. In doing so, it builds on the information presented in a companion handbook in this series (*How to Review the Evidence: Systematic Identification and Review of the Scientific Literature*), drawing on the most recent methods that have emerged in this rapidly developing area.

The way in which the guidance provided in this handbook fits into the overall guideline development process recommended by the NHMRC is shown in the flowchart on page ix. Other handbooks that have been produced in this series so far are:

How to Review the Evidence: Systematic Identification and Review of the Scientific Literature

How to Put the Evidence into Practice: Implementation and Dissemination Strategies

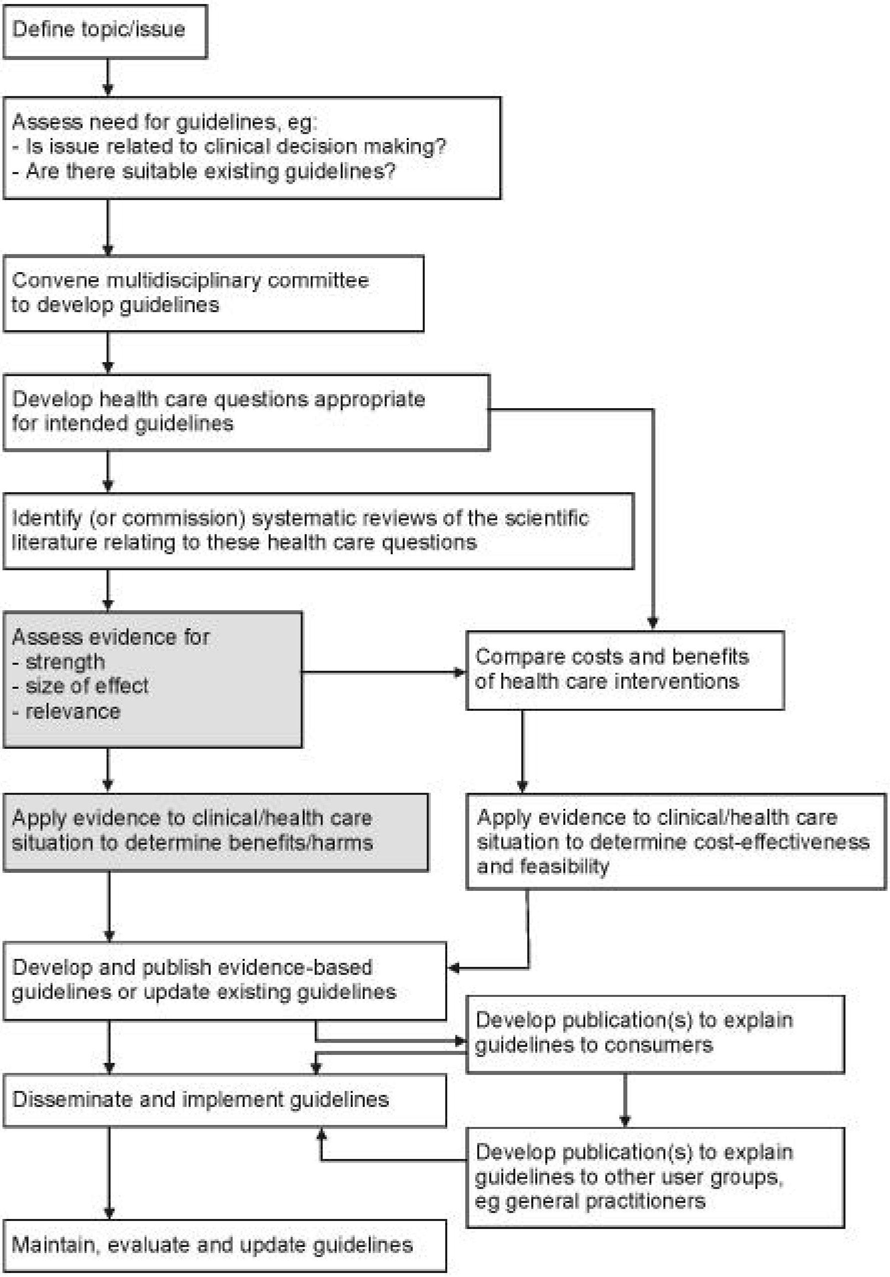
How to Present the Evidence for Consumers: Preparation of Consumer Publications

How to Compare the Costs and Benefits: Evaluation of the Economic Evidence

The series may be expanded in the future to include handbooks about other aspects of the guideline development process, as well as related issues such as reviewing and evaluating evidence for public health issues.

Flow chart showing the clinical practice guidelines development process

(Shaded boxes show the stages described in this handbook)


Define topic/issue
Assess need for guidelines, eg:
- Is issue related to clinical decision making?
- Are there suitable existing guidelines?
Convene multidisciplinary committee to develop guidelines
Develop health care questions appropriate for intended guidelines
Identify (or commission) systematic reviews of the scientific literature relating to these health care questions
Assess evidence for
- strength
- size of effect
- relevance
Compare costs and benefits of health care interventions
Apply evidence to clinical/health care situation to determine benefits/harms
Apply evidence to clinical/health care situation to determine cost-effectiveness and feasibility
Develop and publish evidence-based guidelines or update existing guidelines
Develop publication(s) to explain guidelines to consumers
Disseminate and implement guidelines
Develop publication(s) to explain guidelines to other user groups, eg general practitioners
Maintain, evaluate and update guidelines

INTRODUCTION

Development of evidence-based guidelines

The process for clinical practice guideline development is described in the National Health and Medical Research Council (NHMRC) publication *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999). This recommends that guidelines should be developed by a multidisciplinary guideline development committee, the initial task of which is to determine the need for and scope of the guidelines, define the purpose and target audience and identify the health outcomes that will improve as a result of their implementation.

The membership of a guideline development committee will depend on the nature of the particular guidelines being developed but will include clinicians, health professionals, consumers, health policy analysts, economists and regulatory agency representatives, industry representatives and bioethicists (see NHMRC 1999 for a full list and further discussion of the multidisciplinary committee). The inclusion of consumers is particularly important to ensure that patient-relevant outcomes and evidence are considered.

At the core of an evidence-based approach to clinical or public health issues is inevitably the evidence itself, which needs to be carefully gathered and collated from a systematic literature review of the particular issues in question (see *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature* in this series, NHMRC 2000a). Interpretation of this evidence and its use to frame appropriate guidelines or recommendations has been a major challenge for expert committees compiling clinical practice guidelines over the last few years as an evidence-based approach has been developed and trialed. Because the interpretation of evidence can be a difficult task, the guideline development committee may need to seek the advice of a biostatistician.

Using and presenting evidence

The type of evidence required will vary with the clinical question. *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature* (NHMRC 2000a) in this series lists six different types of questions.

- *Interventions* ('What are the effects of an intervention?')
  ideal study design is a randomised controlled trial (RCT).

- *Frequency or rate* ('How common is a particular condition or disease in a specified group in the population?') — for which the ideal study design is a cross-sectional survey with a standardised measurement in a representative (eg random) sample of people; for a rate, the sample population would need to be followed over time.

- *Diagnostic test performance* ('How accurate is a sign, symptom, or diagnostic test in predicting the true diagnostic category of a patient?') — for which the ideal study design is a representative sample of people in whom the new test is compared to an appropriate 'gold standard' (case-control or cross-sectional study).

- *Aetiology and risk factors* ('Are there known factors that increase the risk of the disease?') — for which the ideal study design is long-term follow-up of a representative inception cohort or an approximation to this through a case-control study.

- *Prediction and prognosis* ('Can the risk for a patient be predicted?') — for which the ideal study design is long-term follow-up of a representative inception cohort (cohort or survival study).

- *Economic analysis* ('What are the overall costs of using the procedure?') — which is not a separate question but involves additional outcomes (resources and their costs) within one of the other questions.The issues of economic evaluation and cost-effectiveness are discussed briefly in the handbook on *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature* (NHMRC 2000a) and in greater detail in the handbook on *How to Compare the Costs and Benefits: Evaluation of the Economic Evidence* (NHMRC 2000b).

In this current guide, we focus on the question of using the evidence about interventions. While the other clinical questions are of equal importance, the advice about how to use the evidence for these is less well developed and so will not be covered here. An intervention will generally be a therapeutic procedure such as treatment with a pharmaceutical agent, surgery, a dietary supplement, a dietary change or psychotherapy. Some other interventions are less obvious, such as early detection (screening), patient educational materials, or legislation. The key characteristic is that a person or their environment is manipulated in order to benefit that person.

The types of evidence that underpin advice on the value of different interventions will vary. It is reasonable to expect that the evidence regarding the efficacy of most drugs will be in the form of RCTs, or (increasingly) systematic reviews of RCTs. However, information on the efficacy of surgical procedures, the clinical impact (rather than the accuracy) of diagnostic tests, and the adverse effects of a range of health interventions, will usually come from nonrandomised studies. The same will often be true for public health interventions.

While it is important to recognise the possibility of bias when the evidence for interventions comes from nonrandomised studies, opportunities should not be missed to improve public health or clinical care even if RCTs have not, or cannot, be performed. For this reason, the use of a single measure to rate the evidence (eg ‘a level A recommendation’) is not favoured by the Health Advisory Committee of the NHMRC. Instead, the NHMRC 1999 guide recommends that all dimensions of the available evidence be assessed in light of what is *feasible and appropriate* (in relation to the intervention(s) under review). It is suggested that these considerations are presented as a brief evidence summary to accompany the main recommendations. This handbook concentrates on issues of applicability of evidence about interventions related to the prevention and treatment of diseases or other outcomes. Methods for assessing the applicability of the results of evaluations of the accuracy of diagnostic tests are not sufficiently well developed to be included at this stage. However, as methodological work is completed, the handbook should be revised to incorporate the most up-to-date information.

About this handbook

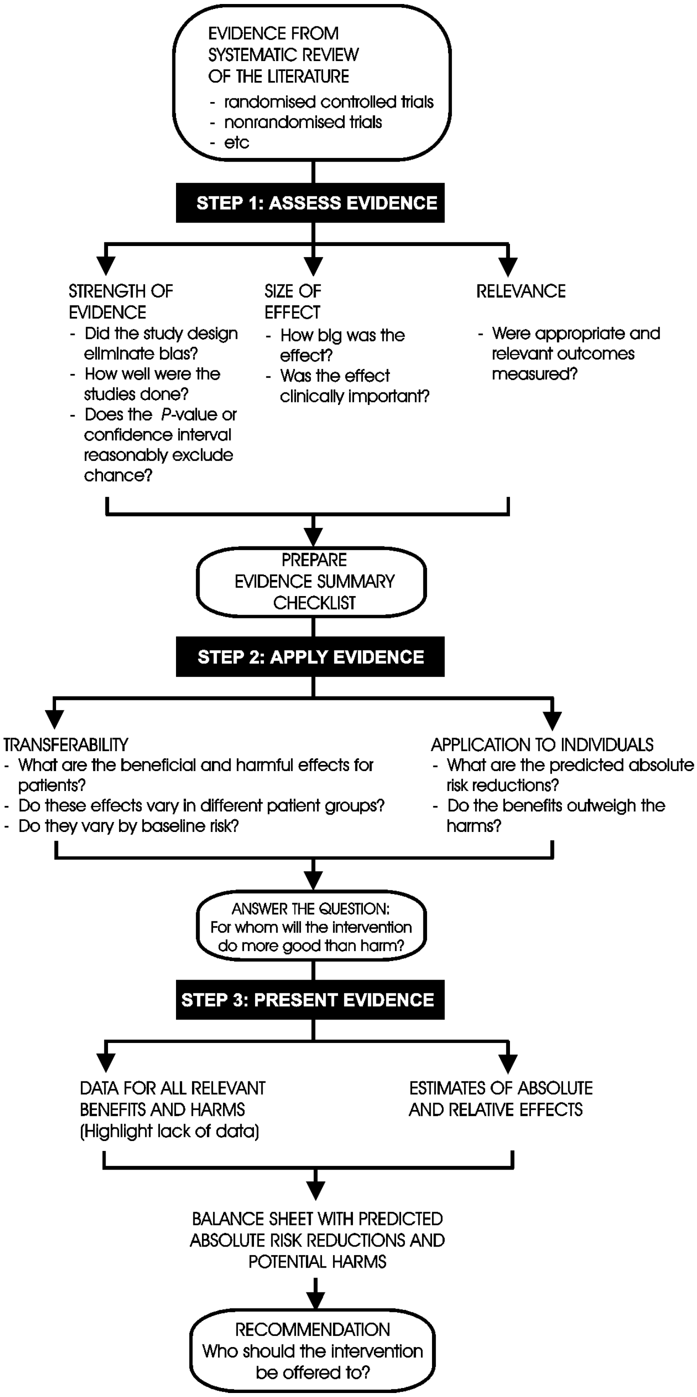
Wherever possible, the recommendations made in this handbook are evidence based. The group contributing to the handbook has undertaken previous systematic reviews on the following topics: methods for the assessment of health technologies; methods for assessing the generalisability of findings from RCTs and systematic reviews; and the effects of information framing (the format in which estimates of treatment effect are presented) on clinical decision making. Where specific recommendations are made in this handbook, they are supported with a brief summary of the key issues and evidence.

This handbook is divided into three sections.

1. How to assess the evidence in terms of the important dimensions, which are defined below.
2. How to apply evidence so that treatment recommendations maximise the probability of benefit and minimise the probability of harm.
3. How to present the data on the benefits and harms of the intervention(s) to health professionals and consumers.

These steps are summarised in the flowchart below.

Steps for assessing and applying scientific evidence



EVIDENCE FROM SYSTEMATIC REVIEW OF THE LITERATURE
- randomised controlled trials
- nonrandomised trials
- etc

STEP 1: ASSESS EVIDENCE

STRENGTH OF EVIDENCE
- Did the study design eliminate bias?
- How well were the studies done?
- Does the *P*-value or confidence interval reasonably exclude chance?

SIZE OF EFFECT
- How big was the effect?
- Was the effect clinically important?

RELEVANCE
- Were appropriate and relevant outcomes measured?

PREPARE EVIDENCE SUMMARY CHECKLIST

STEP 2: APPLY EVIDENCE

TRANSFERABILITY
- What are the beneficial and harmful effects for patients?
- Do these effects vary in different patient groups?
- Do they vary by baseline risk?

APPLICATION TO INDIVIDUALS
- What are the predicted absolute risk reductions?
- Do the benefits outweigh the harms?

ANSWER THE QUESTION: For whom will the intervention do more good than harm?

STEP 3: PRESENT EVIDENCE

DATA FOR ALL RELEVANT BENEFITS AND HARMS (Highlight lack of data)

ESTIMATES OF ABSOLUTE AND RELATIVE EFFECTS

BALANCE SHEET WITH PREDICTED ABSOLUTE RISK REDUCTIONS AND POTENTIAL HARMS

RECOMMENDATION Who should the intervention be offered to?

1 ASSESSING THE EVIDENCE

The thinking about assessing the evidence has evolved since the publication of *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999a). Thus readers will find that the definitions of the dimensions of evidence used in this handbook differ from those outlined in pages 14 to 17 and Appendix A of that publication.

In particular, the definition of *strength of evidence* has changed.

1.1 Dimensions of evidence

When reviewing the evidence three questions regarding the intervention should be considered:

1. Is there a real effect?
2. Is the size of the effect clinically important?
3. Is the evidence relevant to practice?

The first question relates to whether the treatment effect could be due to bias (ie the level and quality of the evidence) or chance (the statistical precision or *P*-value). The size (or magnitude) of the effect should be compared with an effect that is likely to make a clinically important difference.

These terms are defined in Table 1.1. There have been a number of suggestions to roll some of these dimensions into a single classification. While this, in one respect, is appealing (as it simplifies the process and provides a single classification), a lot of valuable information about the evidence can be lost. However, the first three dimensions (level, quality and statistical precision) collectively are a measure of the strength of the evidence. Each of the dimensions shown in Table 1.1 are discussed in detail in the following Sections1.2 –1.6.

1.2 Level of evidence

The *level* of evidence indicates the study design used by the investigators to assess the effectiveness of an intervention. The level assigned to a study reflects the degree to which bias has been eliminated by the study design.

Table 1.1 Evidence dimensions — definitions

Type of evidence (dimension)	Definition
Strength of evidence	
Level	The study design used, as an indicator of the degree to which bias has been eliminated by design (see Table 1.3).
Quality	The methods used by investigators to minimise bias within a study design.
Statistical precision	The *P*-value or, alternatively, the precision of the estimate of the effect (as indicated by the confidence interval). It reflects the degree of certainty about the existence of a true effect.
Size of effect	The distance of the study estimate from the 'null' value and the inclusion of only clinically important effects in the confidence interval.
Relevance of evidence	The usefulness of the evidence in clinical practice, particularly the appropriateness of the outcome measures used.

1.2.1 Study design

The types of studies that are most commonly used to assess clinical and public health issues are shown in Table 1.2.

The study designs shown in Table 1.2 have been used as the basis of the levels of evidence for interventions that are included in *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999), as shown in Table 1.3. These levels of evidence are a convenient way of summarising study designs according to their generally perceived capacity to minimise or eliminate bias in the effect being measured. It is important that they are not perceived to represent the *strength of evidence*, to which study design is only one of several contributors.

Table 1.2 Types of studies used for assessing clinical and public health interventions

Study design	Protocol
Systematic review	Systematic location, appraisal and synthesis of evidence from scientific studies.
Experimental studies	
Randomised controlled trial	Subjects are randomly allocated to groups for either the intervention/treatment being studied or control/placebo (using a random mechanism, such as coin toss, random number table, or computer-generated random numbers) and the outcomes are compared.
Pseudorandomised controlled trial	Subjects are allocated to groups for intervention/treatment or control/placebo using a nonrandom method (such as alternate allocation, allocation by days of the week, or odd–even study numbers) and the outcomes are compared.
Clustered randomised trial	Groups of subjects are randomised to intervention or control groups (eg community intervention trials).
Comparative (nonrandomised and observational) studies	
Concurrent control or cohort	Outcomes are compared for a group receiving the treatment/intervention being studied, concurrently with control subjects receiving the comparison treatment/intervention (eg usual or no care).
Case-control	Subjects with the outcome or disease and an appropriate group of controls without the outcome or disease are selected and information is obtained about the previous exposure to the treatment/intervention or other factor being studied.
Historical control	Outcomes for a prospectively collected group of subjects exposed to the new treatment/intervention are compared with either a previously published series or previously treated subjects at the same institutions.
Interrupted time series	Trends in the outcome or disease are compared over multiple time points before and after the introduction of the treatment/intervention or other factor being studied.

... contd

Table 1.2 (contd)

Study design	Protocol
Other observational studies	
Case series	A single group of subjects are exposed to the treatment/intervention.
– post-test	Only outcomes after the intervention are recorded in the case series, so no comparisons can be made.
– pretest/post-test	Outcomes are measured in subjects before and after exposure to the treatment/intervention for comparison (also called a 'before-and-after' study).

Table 1.3 Designation of levels of evidence

Level of evidence	Study design
I	Evidence obtained from a systematic review of all relevant randomised controlled trials.
II	Evidence obtained from at least one properly-designed randomised controlled trial.
III-1	Evidence obtained from well-designed pseudorandomised controlled trials (alternate allocation or some other method).
III-2	Evidence obtained from comparative studies (including systematic reviews of such studies) with concurrent controls and allocation not randomised, cohort studies, case-control studies, or interrupted time series with a control group.
III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group.
IV	Evidence obtained from case series, either post-test or pretest/post-test.

Source: NHMRC 1999

Systematic review

A systematic review of RCTs is considered generally to be the best study design for assessing the effect of interventions because it has (or should have) identified and examined *all* of the randomised controlled trial evidence about the intervention or treatment. Through the examination of sources of heterogeneity and similarities and differences between the trials, a good understanding of the magnitude of treatment effect and how it varies across relevant factors can be obtained.

Randomised controlled trials

Individually, well-designed and conducted RCTs are the best source of evidence for effects of interventions because randomisation minimises biases that may occur when individuals are allocated in an open fashion to the intervention or control groups. It also minimises or eliminates confounding due to an unequal distribution between the groups, of factors that influence the clinical outcome under study.

Pseudorandomised controlled trials

Pseudorandomised controlled trials are also experimental studies (in the sense that the investigator determines who will get the intervention without regard to the characteristics of the individual eligible subjects) but the main concern is that, because the allocation procedure is known to all, the inclusion of subjects in the trial may be influenced by knowledge of the treatment group to which they are to be allocated. This may introduce biases, including confounding, in the treatment comparisons and thus compromise the outcome of the study.

Comparative (nonrandomised and observational) studies

There are a number of potential problems with nonrandomised studies (which include comparative studies with concurrent or historical control groups, cohort studies and case-control studies). These include:

- noncomparability of groups due to purposive selection of subjects to receive the intervention;

- different cointerventions and other medical management being received by the groups being compared; and

- different methods of outcome measurement being used in each of the groups being compared.

Of these the first is probably the most important and insurmountable within nonrandomised studies; for example, selection bias where those allocated to receive the new treatment or intervention are chosen (consciously or unconsciously) because they are expected to do well. There is also the ‘healthy cohort effect’ in which individuals who are inherently healthier, or more compliant, self-select to the intervention of interest (eg by choosing to take hormone replacement therapy after menopause) (Barrett-Connor 1991).

The use of historical controls (a group who received what was, in the past, standard care) to compare with a group who have more recently received the new treatment is also problematic. The new treatment may have been offered selectively to patients in whom it is likely to succeed. Also, a number of other factors (such as cointerventions, medical management, outcome determination) in addition to the introduction of the new treatment, may explain the observed differences between the two groups.

The comparative studies corresponding to the lower levels of evidence in Table 1.3 have significant potential for bias. Comparison of two groups of subjects from different RCTs (eg to compare two active treatments each of which have been compared with placebo in different trials) is observational in nature. The protection against confounding afforded by the randomisation in the original trials is lost in this type of comparison and there may be a number of confounding factors, including time, place, setting and patient characteristics, which bias the estimate of treatment effect.

When there is no control group, evaluation of the treatment effect must be based on a comparison of outcomes (eg signs, symptoms, severity of disease), in the same group of patients, before and after receiving the treatment. These types of studies are referred to in Tables 1.2 and 1.3 as interrupted time series studies (when the outcomes are measured at multiple time points both before and after the intervention) or a case series with a pretest and post-test. The problem with these designs is the difficulty in attributing any observed changes to the intervention. For example, an analysis of the effect of feeding back prescribing data to general practitioners suggested a decrease in prescribing of some drug groups after the intervention. However, an identical trend was observed in the control (no feedback) group (O’Connell et al 1999).

Expert opinion

Unlike an earlier version of the levels of evidence (NHMRC 1995), the current levels (see Table 1.3) exclude expert opinion and consensus from an expert committee as they do not arise directly from scientific investigation. A comparison of the evidence accumulating over time on the effectiveness of treatments for myocardial infarction (using a method called cumulative meta-analysis) with recommendations made by clinical experts in textbooks and review articles, found that often the expert opinions were not in line with the evidence —either failing to recommend treatments that were effective or recommending routine use of ineffective or potentially harmful treatments (Antman et al 1992). However, when evidence is lacking, this form of recommendation may be the only option. In this case it should be acknowledged that the recommendation is based on expert opinion, rather than evidence, and it should be reviewed as new evidence becomes available.

1.2.2 What if studies disagree?

The results of randomised and nonrandomised evaluations of an intervention can lead to different conclusions about the treatment's effectiveness. For example, nonrandomised (cohort) studies of the effects of antioxidants, such as beta-carotene, in the prevention of various cancers (Byers and Perry 1992) and of the effect of hormone replacement therapy in preventing coronary heart disease (Grady et al 1992) suggested that there were strong protective effects, while the corresponding RCTs found no effect (TCCPSG 1994, Omenn et al 1996, Hennekens et al 1996, Hulley et al 1998). In these examples a 'healthy cohort effect' or lack of control of confounding by other factors may have introduced biases in the nonrandomised studies that resulted in an apparent benefit that was not confirmed in more rigorous studies. In this situation, there are strong arguments for accepting the results of the randomised trials as providing more accurate estimates of the true treatment effect, at least in the conditions under which the treatment was given in these studies.

However, there are situations where the results of nonrandomised and randomised studies are in agreement. For example, the effectiveness of preoperative autologous blood donation in reducing the need for homologous blood transfusion has been assessed in both cohort studies and RCTs. The reduction in the odds of requiring an homologous blood transfusion was similar from a pooling of six RCTs (odds ratio [OR] 0.17, 95% confidence interval [CI] 0.08 to 0.32) and of nine cohort (nonrandomised) studies (OR 0.19, 95% CI 0.14 to 0.26) (Forgie et al 1998).

Sometimes there is apparent disagreement between randomised trials. For instance, the early trials of the drugs dipyridamole and aspirin concluded that this combination was no more effective than aspirin alone in preventing thromboembolic stroke in high-risk individuals (ACCSG 1983, Bousser et al 1983). A recent large trial found a clinically and statistically significant advantage of the combination therapy (Diener et al 1996). In this case, pooling of the data shows that there is no true disagreement between the results of the old and new trials (there was no significant heterogeneity).

The most likely explanation for the apparent discrepancy is that the older trials were too small to detect the effect of the combination treatment. This is an example of failure to adequately weigh the possible effects of chance in producing the apparent lack of benefit of the combination therapy. This is a common explanation for apparent disagreements between trials, and one that is easily identified by meta-analysis techniques. Analyses within a systematic review can be used to explore the reasons for true variation between trial results and these are discussed later in this handbook.

There has been considerable interest in the disagreements that sometimes arise between the results of large RCTs and systematic reviews. This has led some authors to question the accuracy of the systematic reviews, particularly if they involved small trials. However, careful examination of the data sometimes provides plausible explanations for these discrepancies.

In a systematic review of RCTs of thrombolytic therapy in acute ischaemic stroke, Wardlaw et al (1997) attempted to explain the conflicting results seen in the individual trials. The heterogeneity in the trial results could be explained in part by differences in the use of antithrombotic drugs within 24 hours of thrombolysis, different levels of stroke severity, differences in the delay between occurrence of the stroke and thrombolysis and differences in dose of thrombolytic drug used.

In other instances possible explanations for apparent differences between systematic reviews and large trials rest on the results of quite sophisticated analytical techniques. The example of the administration of intravenous magnesium following myocardial infarction is described in Box 1.1 below.

1.2.3 Can causation be established in the absence of randomised controlled trials?

The levels of evidence based on study design may appear to rule out valid causal inferences being made in the absence of randomised controlled trials. If true, this would be problematic for evaluation of some public health and health services interventions where randomisation is difficult.

However, the Canadian Advisory Committee on Community Health convened the Community Health Practice Guidelines Working Group to develop an approach for formulating evidence-based recommendations regarding community health interventions (including health promotion programs, communicable disease control and environmental health control) (Gyorkos et al 1994). This working group recognised that the considerable variability in the nature and extent of the available evidence and the diversity in the interventions to be evaluated, made it difficult to assess the evidence. However, they argued that the process should be (and could be) adapted accordingly. They applied their approach to three high priority community health interventions:

- immunisation delivery methods;
- partner notification strategies for sexually transmitted diseases; and
- restaurant inspection programs.

The evidence reviewed for each intervention was restricted to relevant comparative studies (including pretest and post-test studies) so that the strength of the association between the intervention and outcome could be assessed. Careful consideration was given to the internal validity of each study using established quality criteria and each study was classified as strong (no major flaws and a few minor flaws), moderate (no major flaws but some minor ones) or weak (one or more major flaws). This demonstrates that the lack of level I or level II evidence does not preclude a rigorous approach to the assessment of the available evidence when developing practice guidelines.

Box 1.1 Comparison of the results from a systematic review and a single large trial

Problem

A systematic review of the intravenous administration of magnesium after myocardial infarction (Yusuf et al 1993) showed a significant benefit for the intervention.
A single large randomised controlled trial (ISIS-4 1995) showed no benefit.

Response

Some authors were quick to dismiss the results of the systematic review as misleading. However, others attempted to explore the possible reasons for the disagreement. The following valid suggestions appeared in the literature to explain the difference:

- publication bias (when the results of a trial are more likely to be published if a statistically significant benefit of treatment was found) may account for the results of the systematic review (Egger and Davey-Smith 1995);
- the smaller, earlier trials on which the systematic review is based included higher-risk individuals while the large (or mega) trial (which had more relaxed inclusion criteria) included lower-risk individuals (Cappelleri et al 1996).

Cappelleri et al (1996) also demonstrated that if there is a relationship between the magnitude of the treatment effect and baseline (or untreated) risk, then the small trials would be expected to find a larger treatment effect compared with the megatrial and this could explain the discrepancy. They found that for magnesium after myocardial infarction, there is an increase in relative risk reduction with increasing baseline risk and that the baseline risk for subjects included in ISIS-4 was lower than the average baseline risk across the smaller studies.

Conclusion

If study results do not agree, it is useful to look for an explanation. The approach of Cappelleri et al 1996 in this case was particularly informative in terms of the applicability of the results (see Section 2).

1.2.4 Need for different study designs

It may be necessary to use different study designs for different aspects of the intervention’s effect. For example, there may be level I evidence on the potential benefits of the treatment. However, the individual trials may have been too small, or of insufficient duration to provide good estimates of the potential harmful effects of treatment so that the best data about these outcomes will come from cohort studies or case-control studies based on use of the intervention in wider practice.

For example, the best information about the harmful side effects of the anticoagulant drug warfarin (ie increased risk of major bleeding) comes from a cohort study. In five RCTs of the use of warfarin to prevent stroke in individuals with nonrheumatic atrial fibrillation, the increase in absolute risk of major bleeding was 0.6% (an increase from 0.8% to 1.4% per year). This compares with an excess risk of around 3% estimated from population-based observational studies (Glasziou et al 1998). Also, case-control studies have been the primary source of data on the risk of gastrointestinal bleeding associated with use of nonsteroidal anti-inflammatory drugs (NSAIDs) (Henry et al 1996).

Key points for considering levels of evidence

1. Differences in the conclusions reached about effectiveness from studies at differing levels of evidence or within a given level of evidence need to be resolved.

2. Resolving these discrepancies should be viewed as an important task in the compilation of an evidence summary.

3. Biostatistical and epidemiological advice may be needed on how to search for possible explanations for the disagreements before data are rejected as being an unsuitable basis on which to make recommendations.

4. It may not be feasible to undertake an RCT in all situations. But, regardless of the clinical context, guidelines should be based on the best available evidence and if this evidence is suboptimal (eg based on observational data because an RCT, although feasible, has not been done) then this should be acknowledged.

5. It may be necessary to use evidence from different study designs for different aspects of the treatment effect. In general, there should be studies providing higher level evidence on the benefits.

1.3 Quality of the evidence

The *quality* of the evidence refers to the methods used by the investigators during the study to minimise bias and control confounding within a study type (ie how well the investigators conducted the study).

The important sources of bias, and their possible effects, depend on the type of study. Various quality assessment instruments have been proposed — the most well-developed and tested are those for RCTs (Moher et al 1996). Some authors have suggested quality criteria for nonrandomised studies (including cohort and case-control studies) but their validity and reliability are largely untested (Gyorkos et al 1994, Powell et al 1987, Realini and Goldzieher 1985, Carson et al 1994, Lichtenstein et al 1987, Horwitz and Feinstein 1979).

If the main source of evidence is a systematic review of RCTs, the quality of the review should be assessed. Standard methods for conducting and reporting systematic reviews have been reported by a number of authors (Greenhalgh 1997, Hunt and McKibbon 1997).

The quality issues that appear to be important for each study type are summarised in Table 1.4. Examples of quality assessment instruments are provided in the accompanying handbook in this series on *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature* (NHMRC 2000a).

1.3.1 Is study quality important?

A number of studies have examined the effect of study quality on the estimates of treatment effect. In general, RCTs that were not double blind (ie the subjects and the observers were not ignorant of the treatment group) and in which concealment of allocation was inadequate, tend to result in larger estimates of treatment effect (in favour of the new intervention) (Schulz et al 1995, Colditz et al 1989, Moher et al 1998). For studies that were unclear about whether treatment allocation was adequately concealed, the OR for treatment effect was, on average, 30% lower (in favour of the new treatment) than for those with adequate allocation concealment (Schulz et al 1995). Also, lack of double blinding exaggerated the estimated treatment effect by an average of 17% on average. A relationship between study quality and benefit from physical activity in reducing coronary heart disease risk was also demonstrated in a systematic review of cohort studies (Berlin and Colditz 1990).

Table 1.4 Quality criteria for randomised controlled trials, cohort studies, case-control studies and systematic reviews

Study type	Quality criteria
Randomised controlled trials[a]	Was the study double blinded? Was allocation to treatment groups concealed from those responsible for recruiting the subjects? Were all randomised participants included in the analysis?
Cohort studies[b]	How were subjects selected for the 'new intervention'? How were subjects selected for the comparison or control group? Does the study adequately control for demographic characteristics, clinical features and other potential confounding variables in the design or analysis? Was the measurement of outcomes unbiased (ie blinded to treatment group and comparable across groups)? Was follow-up long enough for outcomes to occur? Was follow-up complete and were there exclusions from the analysis?
Case-control studies[b]	How were cases defined and selected? How were controls defined and selected? Does the study adequately control for demographic characteristics and important potential confounders in the design or analysis? Was measurement of exposure to the factor of interest (eg the new intervention) adequate and kept blinded to case/control status? Were all selected subjects included in the analysis?
Systematic reviews[c]	Was an adequate search strategy used? Were the inclusion criteria appropriate and applied in an unbiased way? Was a quality assessment of included studies undertaken? Were the characteristics and results of the individual studies appropriately summarised? Were the methods for pooling the data appropriate? Were sources of heterogeneity explored?

[a] Based on work of Schulz et al (1995) and Jadad et al (1996)

[b] Based on quality assessment instruments developed and being tested in Australia and Canada

[c] Based on articles by Greenhalgh (1997) and Hunt and McKibbon (1997)

Key points for considering the quality of the evidence

1. It is important to identify the strengths and weaknesses of the studies that are providing the evidence and to consider the potential biases and their possible influence on the estimate of effect (in particular whether the biases are likely to shift the estimated treatment effect toward, or away from, the null value).

2. Given the potential effects of study quality on the estimates of effect, quality assessment of the individual studies should be undertaken.

3. An assessment of study quality may be helpful in resolving apparent conflicts in conclusions drawn about effectiveness from different studies.

1.4 Statistical precision of the effect

The magnitude of the *P*-value and the precision (or width of the confidence interval) of the estimate of the treatment effect are important when assessing the strength of the evidence. While the two are interrelated, the actual value of *P* (not just a specification that it is less than some arbitrary value) conveys more information than the confidence interval alone.

The *P*-value is often misunderstood. It does not, as commonly believed, represent the probability that the null hypothesis (that there is no treatment effect) is true (a small *P*-value therefore being desirable). The *P*-value obtained from a statistical test corresponds to the probability of claiming that there is a treatment effect when in fact there is no real effect. Incorrect rejection of the null hypothesis is called a Type I error.

It is sometimes helpful to compare the process of hypothesis testing with the evaluation of a diagnostic test against a 'gold standard'. In this analogy, the *P*-value is equivalent to the false-positive rate (Diamond and Forrester 1983).

The *P*-value provides an indication of whether chance alone may explain the treatment effect that has been observed. While there is no 'magical' value of *P*, the smaller *P* is, the greater the certainty that the effect is real.

1.5 Size of the effect

The measure of treatment effect indicates how the new treatment compares with doing nothing, using a placebo or using an alternative active treatment.

The *size* of the treatment effect refers to the size (or the distance from the null value) of the measure (or point estimate) of treatment effect and the values included in the corresponding 95% CI. In the case of a systematic review, it is the summary measure of effect based on the studies included in the review.

1.5.1 Measures of treatment effect

The choice of the measure of treatment effect depends on a number of things, the most important being the scale of measurement used. The effect of some outcomes will be measured on a continuous scale (eg blood pressure, serum cholesterol, quality of life), while others are on a dichotomous or binary scale (eg improved/not improved, dead/alive, cancer recurrence/no recurrence). The most common measures of treatment effect are defined in Table 1.5.

The risk difference cannot be estimated from case-control studies and the relative risk can only be estimated indirectly as the odds ratio.

Many people find the odds ratio difficult to interpret in a clinical setting. However, it is often used in the analysis of RCTs and cohort studies because it has desirable statistical properties. For ease of clinical interpretation, if odds ratios are used in the assessment of the treatment effect, they should be converted back to relative risks or risk differences for data presentation. This may require collaboration with a biostatistician.

While the relative risk is appropriate for the assessment of the overall treatment effect, it is difficult to interpret for an individual. The risk difference is of more interest to an individual patient as, for a given relative risk reduction (say 50%), it is important to know whether it represents a reduction in risk from 10% to 5% or from 2 per million to 1 per million. Many people find the number needed to treat (NNT) easier to understand.

Table 1.6 shows the risk differences and NNTs corresponding to different relative risk reductions and levels of baseline risk. While the relative risk is constant, the absolute risk reduction decreases and NNTs increase with decreasing levels of baseline risk. See Section 2 for further details.

Table 1.5 Measures of treatment effect for continuous and binary outcomes

Outcome measure	Description
Continuous outcomes	
Difference between group means	Difference between treatment and control groups in mean values of outcome variable.
Standardised difference	Differences between the treatment and control group means for each study, standardised by an estimate of the standard deviation of the measurements in that study. This removes the effect of the scale of measurement, but can be difficult to interpret.
Weighted difference in means	Average (pooled) difference between treatment and control groups in mean values across a group of studies using the same scale of measurement for the outcome (eg blood pressure measured in mm Hg).
Standardised weighted mean difference	Average (pooled) standardised difference between treatment and control groups across a group of studies, where the outcome was measured using different scales with no natural conversion to a common measure (eg different depression scales or different quality-of-life instruments).
Binary outcomes	
Risk difference (RD)	Difference (absolute) between treatment and control groups in the proportions with the outcome. If the outcome represents an adverse event (such as death) and the risk difference is negative (below zero) this suggests that the treatment reduces the risk. In this situation the risk difference, without the negative sign, is called the *absolute risk reduction*.
Relative risk or risk ratio (RR)	Ratio of the proportions in the treatment and control groups with the outcome. This expresses the risk of the outcome in the treatment group relative to that in the control group. For an adverse outcome, if the relative risk is below 1, this suggests that the treatment reduces the risk; its complement (1–relative risk) or *relative risk reduction* is also often used.
Odds ratio (OR)	Ratio of the odds of the outcome in the treatment group to the corresponding odds in the control group. Again, for an adverse outcome, an odds ratio below 1 indicates that the treatment reduces the risk. In some studies (eg population-based case-control studies) the odds ratio is a reasonable estimate of the relative risk. It is not a good estimate when the outcome is common or is measured as a prevalence.
Hazard ratio (HR)	Ratio of the hazards in the treatment and control groups (when time to the outcome of interest is known); where the hazard is the probability of having the outcome at time *t*, given that the outcome has not occurred up to time *t*. Sometimes, the hazard ratio is referred to as the *relative risk*. For an adverse outcome, a hazard ratio less than 1 indicates that the treatment reduces the risk of that outcome.
Number needed to treat (NNT)	The number of patients who have to be treated to prevent one event. It is calculated as the inverse of the risk difference without the negative sign (NNT = 1/RD). When the treatment increases the risk of the harmful outcome, then the inverse of the risk difference is called the number needed to harm (NNH = 1/RD).

Table 1.6 Absolute reductions in risk associated with relative risk reductions of 50% and 25%

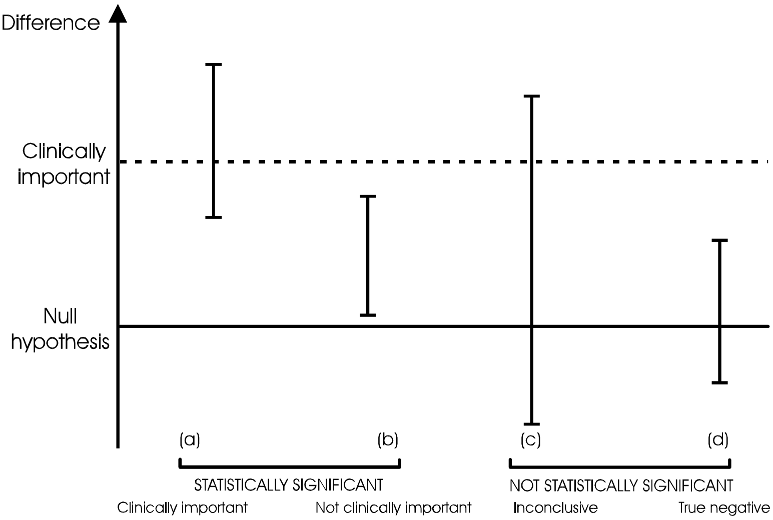
	50% Relative risk reduction (RR=0.50)		25% Relative risk reduction (RR=0.75)	
Baseline risk	**Risk difference**	**NNT**	**Risk difference**	**NNT**
10%	5%	20	2.5%	40
1%	0.5%	200	0.25%	400
1 in 1000	0.05%	2000	0.025%	4000
1 in 10,000	1 in 20,000	20,000	1 in 40,000	40,000

RR=relative risk; NNT=number needed to treat

1.5.2 Interpretation of statistical significance and the size of effect

In large trials, statistically significant effects may be too small to be clinically important. Thus the size of the effect is important as well as its statistical significance. For the evidence to suggest that an intervention is useful, the 95% CI should include only clinically important treatment effects. Figure 1.1 shows how the position of the 95% CI (relative to the null value and a difference that is considered to be clinically important) and its size (or width) indicates the intervention's effect in terms of statistical significance and clinical importance (Berry 1986).

Ideally, the recommendations made in clinical practice guidelines will be based on effects that are both highly statistically significant *and* clinically important so that the 95% CI includes clinically important effects (case (a) in Figure 1.1). The situations illustrated as cases (b) and (d) are also informative results — demonstrating that the intervention is unlikely to produce clinically important effects. Situations like case (c) should be avoided as the 95% CI is consistent with both no treatment effect and clinically important effects. This also illustrates that the lack of statistical significance does not necessarily imply 'no effect'.



Notes: Vertical lines represent the 95% CI around the difference between the treatment and control:
(a) difference statistically significant and clinically important;
(b) difference statistically significant and clinically unimportant;
(c) difference is not statistically significant and of uncertain clinical importance;
(d) difference is not statistically significant and not clinically important.

Figure 1.1 Distinction between statistical significance and clinical importance (Berry 1986)

The two trials described in Box 1.2 illustrate the difference between statistically significant and clinically important treatment effects.

A suggested classification scheme for size of the effect is shown in Table 1.7.

Box 1.2 Statistical precision versus clinically important treatment effects

Trial 1

Comparison of clopidogrel with aspirin for atherosclerotic vascular disease (>19,000 patients; CAPRIE Steering Committee 1996)

Findings
Reduction in the risk of a composite outcome of ischaemic stroke, myocardial infarction or vascular death from 5.83% per year in the aspirin group to 5.32% in the clopidogrel group (*P*=0.043).

Relative risk reduction = 8.7% (95% CI 0.3 to 16.5%).

Statistical significance and clinical importance
While the treatment effect is statistically significant at the 5% level, it is doubtful whether an annual decrease in risk of this outcome of 0.5% (which corresponds to an NNT of 200) with a lower confidence limit of 0.01% (corresponding to an NNT of 10,000) is clinically important.

Mortality from all causes, which may be more important to patients, was not statistically significantly different (3.05% vs 3.11% per year; *P*=0.71).

Trial 2

Intravenous streptokinase versus placebo in patients with suspected heart attack (17,187 patients; ISIS-2 Collaborative Group 1988).

Findings
Vascular deaths during the first five weeks were reduced from 12.0% to 9.2% with a corresponding reduction in odds of death of 25% (95% CI 18 to 32%; *P*<0.00001) in the streptokinase-treated group. This was accompanied by a similar reduction in all-cause mortality (*P*<0.0001).

Statistical significance and clinical importance
This highly statistically significant treatment effect (corresponding to an NNT of 36) has a 95% CI that includes only clinically important effects.

Overall conclusion

The evidence obtained from Trial 1 (the CAPRIE trial) is considered to be weak in terms of statistical precision with a small (and clinically unimportant) treatment effect. In contrast, the results from Trial 2 (the ISIS-2 trial) suggest a highly statistically significant treatment effect of sufficient magnitude to be clinically important.

Table 1.7 Classifying size of the effect

Ranking	Clinical importance of benefit
1	A clinically important benefit for the full range of plausible estimates The confidence limit closest to the measure of no effect (the 'null') rules out a clinically unimportant effect of the intervention
2	The point estimate of effect is clinically important BUT the confidence interval includes clinically unimportant effects
3	The confidence interval does not include any clinically important effects
4	The range of estimates defined by the confidence interval includes clinically important effects BUT the range of estimates defined by the confidence interval is also compatible with no effect, or a harmful effect

Key points for considering the size of the effect

1. The size of the effect is important because it relates to the clinical importance of the effect.

2. The size of the effect should be expressed in both relative and absolute terms (ie as relative risks and absolute risk reductions or NNT for a range of baseline risks).

3. The size of the effect and the certainty with which it is known should both be assessed (see Table 1.7).

1.6 Relevance of the evidence

A very important dimension (perhaps the most important) is the *relevance* of the evidence. The focus of this section will be on the appropriateness of the outcomes. Are they of importance or of interest to the patient? Are they short-term or long-term effects? How well do they relate, in a causal sense, to outcomes of importance to the patient?

Relevance also includes issues like the extent to which the intervention can be replicated in other settings and the applicability of the study findings to other settings and patient groups unlike those in which its efficacy has been tested. These are considered in detail in Section 2 of this handbook.

1.6.1 Appropriateness of the outcomes

Outcomes used to measure the effect of an intervention may be categorised as surrogate, clinical or patient relevant (and these are not mutually exclusive).

Often, the types of outcomes measured and reported in clinical trials are those that are easy to measure and that can be expected to show changes or differences in a relatively short period of time. These may, or may not be, important to the recipients of the intervention. Definitions of the types of outcomes are shown in Table 1.8.

Table 1.8 Definitions of types of outcomes

Outcome	Definition
Surrogate	A laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. Changes induced by a therapy on a surrogate endpoint should be expected to reflect changes in a clinically meaningful endpoint (Temple 1995).
Clinical	Outcomes that tend to be defined on the basis of the disease being studied; for example, survival in cancer, occurrence of vertebral fractures in treatments for osteoporosis, ulcer healing, walking distance or microbiological 'cure' in the treatment of infections.
Patient-relevant	Outcomes that matter to the patient and their carers. They need to be outcomes that patients can experience and that they care about (eg quality of life, return to normal function). Patient-relevant outcomes may also be clinical outcomes or surrogate outcomes that are good predictors (in a causal sense) of outcomes that matter to the patient and their carers.

Temple (1999) defined an intermediate outcome as an endpoint which is a true clinical endpoint (a symptom or measure of function, such as symptoms of hyperglycaemia, angina frequency, or exercise tolerance) but which is not the ultimate endpoint of the disease such as survival, or the rate of serious and irreversible morbid events (eg myocardial infarction or stroke). Intermediate outcomes occur quite late in the causal chain and represent manifestation of disease. Therefore they are clinical rather than surrogate outcomes and they are likely to be patient-relevant.

When are surrogate outcomes useful?

There is a considerable literature that has developed around the role of surrogate outcomes. It has been suggested that to be useful a surrogate outcome should have the following features, but experience suggests that few do.

- It should be a physiological variable.

- It should exhibit an association with a clinical or patient-relevant outcome of interest.

- There should be a biological and pathophysiological basis for believing that the surrogate outcome is a causal determinant of the clinical outcome in the disease being studied and that it is in the chain of causal events believed to be between the intervention and desired clinical or patient-relevant outcome.

The combination of the last two features suggests that the surrogate should be highly predictive of the clinical variable. This has been assumed for a number of surrogates where there has been a statistical association (or correlation) but no causal relationship. Instead, the surrogate has been a risk marker for the disease. For example, xanthomas may develop as a result of an elevated serum cholesterol level. Removal of the xanthoma by surgery will not reduce the cholesterol level, and lowering the cholesterol level may not alter the xanthoma.

Alternatively, the association has not been sufficiently specific to ensure that a change in the surrogate will be followed by a change in the outcome. For example, a change in blood pressure will predict accurately the risk reduction in stroke but not in coronary heart disease (as its causes are multifactorial).

Another problem associated with surrogate outcomes is that their relationship with the clinical outcome may be confounded by other variables. For example, a drug may have other effects that reduce the risk of the clinical outcome (eg heart attack) in addition to the effect measured by the surrogate outcome (eg change in serum cholesterol).

Examples of surrogate outcomes with their corresponding clinical and patient-relevant outcomes are shown in Table 1.9.

The use of surrogate outcomes (CD4 T-lymphocyte counts and plasma viral load) in the assessment of treatments for HIV/AIDS has drawn considerable attention in both the scientific and lay press. Earlier work suggested that CD4 T-lymphocyte counts alone were inadequate indicators of patient-relevant outcomes. More recent work suggests that the combination of the two measures is a better predictor of outcome (progression to AIDS and death) and of response to antiretroviral therapy and treatment failure, than either of them alone (Mellors et al 1997, O’Brien et al 1997).

There are examples where the use of a surrogate outcome has led to inappropriate policy decisions being made (see Box 1.3).

Table 1.9 Examples of surrogate, clinical and patient-relevant outcomes

Disease	Surrogate outcome	Clinical outcome	Patient-relevant outcome
Cardiovascular	blood pressure	stroke, myocardial infarction, survival	clinical outcomes, QoL
HIV/AIDS	CD4 T-lymphyocyte counts	AIDs events, survival	survival, QoL, adverse effects
Fracture	bone mineral density	bone fracture	symptomatic fracture, functional status
Coronary heart disease	blood cholesterol levels	myocardial infarction, survival	myocardial infarction, survival, symptoms (angina), QoL, pain
Otitis media	microbiological 'cure'	clinical cure	deafness, side effects (eg gastrointestinal symptoms)

HIV=human immunodeficiency virus; AIDS=acquired immune deficiency syndrome; QoL=quality of life

Source: extracted from Fleming and DeMets 1996

Box 1.3 Surrogate versus clinical outcomes

Trial 1

A randomised controlled trial of anti-arrhythmic drugs encainide and flecainide (CAPS 1988).

Findings

A significant reduction in premature ventricular contractions on electrocardiograms (ECGs) (surrogate outcome for sudden death after myocardial infarction).

On the basis of this and other trials, these and other similar drugs were approved for registration.

Trial 2

Large randomised controlled trial of encainide and flecainide (CAST 1989).

Findings

A statistically significant increase in mortality associated with use of the drugs (ie clinically important and patient-relevant outcome).

Conclusion

Encainide, flecainide and other drugs should not have been approved based on the surrogate outcome studied (Trial 1). It is important not to settle for a surrogate outcome in studies if it is possible to use a clinical outcome (as was the case in Trial 2).

What are patient-relevant outcomes?

The clinical outcomes measured in studies of treatment effectiveness tend to be those considered to be important by the investigators and clinicians and that are easily measured and generally occur in the shorter term. In the United States, the Agency for Health Care Policy and Research advocates that observable clinical outcomes are of primary importance in the evaluation of interventions (Hadorn and Baker 1994).

However, it is becoming increasingly apparent that this approach does not necessarily capture all of the relevant outcomes from the patient's perspective. In most cases, trials and systematic reviews concentrate on a primary outcome of interest and a limited number of secondary outcomes. These usually reflect potential benefits of the treatment and often the potential harms (side effects and adverse events) are either not measured or not reported even though they may be of primary importance to patients. The outcomes being measured may not reflect how a patient feels, or whether they can go about their usual activities in a satisfactory way. Factors that relate to an improved quality of life and quantity of life (mortality) may be more patient relevant (Hadorn and Baker 1994). These types of measures are indicators of the balance between the treatment's benefits and harms. For example, a drug may significantly reduce blood cholesterol levels (and therefore reduce the risk of coronary heart disease in the future) but result in the patient experiencing decreased quality of life due to the side effects. In particular, for individuals who are at low risk of developing the outcome that the treatment aims to prevent, the predicted absolute risk reduction may be too small to offset the harmful effects of treatment.

A suggested classification of the relevance of the evidence to patients is shown in Table 1.10. In considering the appropriateness of a surrogate outcome measure, this takes into account the setting in which the surrogate has been previously validated. For example, if increased bone mineral density was shown to be a reasonable surrogate for reduced risk of osteoporotic fracture in a trial of one bisphosphonate, then it would be a reasonable surrogate for a trial of another drug in the same class. However, evidence of effect on bone mineral density for another type of drug with a different mode of action (eg hormone replacement therapy) may not be rated as highly.

Table 1.10 Classifying the relevance of the evidence

Ranking	Relevance of the evidence
1	Evidence of an effect on patient-relevant clinical outcomes, including benefits and harms, and quality of life and survival.
2	Evidence of an effect on a surrogate outcome that has been shown to be predictive of patient-relevant outcomes for the same intervention.
3	Evidence of an effect on proven surrogate outcomes but for a different intervention.
4	Evidence of an effect on proven surrogate outcomes but for a different intervention and population.
5	Evidence confined to unproven surrogate outcomes.

Key points for considering patient-relevant outcomes

1. The goal of decision making in health care is to choose the intervention(s) (which may include doing nothing) that is (are) most likely to deliver the outcomes that patients find desirable (Eddy 1990a).

2. Surrogate outcomes (such as blood pressure measurements or levels of serum cholesterol) may be reasonable indicators of whether there has been some effect. However, they should not be the basis for clinical decisions unless they reliably predict an affect on the way the patient feels; otherwise they will not be of interest to the patient or their carers (Eddy 1990a).

3. All possible outcomes that are of most interest to patients (particularly harms) should be identified and evaluated.

4. It may be useful to classify the relevance of the evidence using the categories shown in Table 1.10.

What are inappropriate outcomes?

Any number of examples can be quoted relating to the use of inappropriate outcomes in trials and systematic reviews (ie outcomes that are not directly relevant to a patient's ability to function normally). The outcomes used in trials of treatments for Alzheimer's disease are described here as an illustration (see Box 1.4).

Box 1.4 Inappropriate and appropriate outcome measures for Alzheimer's disease

Inappropriate outcome measures

In RCTs of treatments for Alzheimer's disease, the clinical outcomes measured included:

- measures of cognitive function (including the cognitive subscale of the Alzheimer's Disease Assessment Scale [ADAS-cog];
- the clinicians interview-based impression of change [CIBIC]); and
- the presence or absence of 'clinical improvement' based on the CIBIC scores (Rogers et al 1998a, Knapp et al 1994).

These outcomes may not truly reflect the impact of the disease on the patient and their family. For example, what does a treatment–placebo difference of 2 points on the ADAS-cog subscale (as observed in a trial of tacrine by Knapp et al 1994) actually mean to the patient? Does it translate into an improvement in their ability to undertake the usual activities of daily living?

Appropriate outcome measures

There have been some developments in measuring quality of life (QoL) in both patients and their care givers (Whitehouse 1998). Early work on measuring QoL in patients with dementia used observer ratings of patient behaviour and proxy informants reporting on the patient's QoL. More recently, self ratings from the patient about their QoL have been developed and used in trials of the drug donepezil (Rogers et al 1998b). A recent trial of the drug selegiline went some way towards measuring a more relevant outcome — the time to occurrence of the composite outcome of death, institutionalisation, loss of ability to perform at least two of the three activities of daily living or severe dementia (as defined by a clinical dementia rating of 3) (Sano et al 1997).

1.7 Evidence summary

At this stage, it would be useful to create an overall summary of the evidence gathered so far to form the basis of evidence-based recommendations.

The first step is to summarise the main studies and systematic reviews contributing to the evidence. This should include a description of the intervention, the setting, patient characteristics (including eligibility criteria), outcomes and the main results (in quantitative terms where possible). This could be of a similar format to the table of included studies that appears in all Cochrane Collaboration Reviews and many published systematic reviews and is discussed further in the accompanying handbook on systematically reviewing the scientific literature (NHMRC 2000a).

The next step is to summarise the information about the strength of the evidence (level, quality and statistical precision), the size of the treatment effect and the relevance of the evidence (particularly the appropriateness of the outcomes measured and the omission of any outcomes that are likely to be important to patients).

It cannot be emphasised enough that each of these dimensions should be considered when evaluating and presenting the relevant evidence. Also, the relative importance of the dimensions should be considered in the context of the clinical problem being addressed. For example, evidence from a good quality RCT may be of limited relevance (due to the inappropriate choice of outcomes measured) and so the guideline may need to rely on lower level, but more relevant evidence. The dimensions should also be considered in combination as it may be more appropriate to base a recommendation on a good quality observational study than on a poor quality RCT in which the likelihood of biased results is high.

However, as mentioned previously in this handbook, a reductionist approach whereby all of these dimensions are summarised into a single classification is not recommended as important and useful information will be lost.

A checklist that summarises the data and classifies it according to its level, quality, statistical precision, relevance and the size of the treatment effect should accompany each major recommendation. This checklist should reflect the results, where possible, from a formal synthesis of the available evidence. If there is no systematic review of the relevant studies, the data from the best available studies should be rated.

1.7.1 Suggested format for an evidence checklist

It is difficult to be prescriptive about this as the relative importance of the dimensions will vary according to the treatment and disease under consideration and the evidence at hand. In the final analysis, a judgement will have to be made on whether a recommendation is justified based on the overall adequacy of the available evidence.

A suggested format for the evidence checklist would include a summary of the data about each dimension and a judgment as to whether it is high or low. For example:

Dimension	Score
Strength of evidence	
Level	Level I, II, III, etc (see Table 1.3)
Quality	Score from quality assessment (see Table 1.4)
Statistical precision	*P*-value and width of confidence interval
Size of effect	Summary estimate (eg RR) and 95% confidence interval, plus score for clinical importance of benefit (see Table 1.7).
Relevance of evidence	Score from relevance assessment (see Table 1.10)

1.7.2 Examples

Hypothetical examples of evidence summaries for two clinical recommendations are provided in Box 1.5.

It might be useful to note situations where a dimension is rated as ‘low’, but a higher rating is not achievable. For instance, in the example of NSAIDs in Box 1.5, large RCTs are not feasible so while the evidence scored less well for level, it was the best possible under the circumstances.

1.8 Conclusion

The key issue is that a decision has to be made about what is *feasible and appropriate* in a given situation and the extent to which reasonable standards have been met by the available body of evidence.

The evidence summaries and checklists should provide a clear picture about whether the intervention on balance appears to do more good than harm. The strength of the evidence, the size of the effect and the relevance of the evidence should be such that there is a conviction that the intervention is worth implementing.

Only at this point should the clinical guideline development proceed to the next step, which is to address the question ‘to whom should the intervention be offered so that more good than harm is done?’ This is covered in Section 2 of this handbook.

Box 1.5 Use of evidence — examples

Note: The following examples are hypothetical and do not represent actual recommendations. Information was not available to fully construct the evidence checklist (including quality score, RR, *P*-value and so on), but the general approach is shown.

Example 1 — recommendation to avoid nonsteroidal anti-inflammatory drugs (NSAIDs) in subjects with a history of peptic ulceration

Evidence checklist

Strength	
– level	Case-control studies (level III)
– quality	Good
– statistical precision	Small *P*-values
Size of effect	Large (clinically important) adverse effect of treatment
Relevance	Highly relevant outcome (hospitalisation with major gastrointestinal bleeding).

Conclusion

Although the evidence was obtained from observational studies, the other dimensions rated well and the recommendation can be supported.

Example 2 — recommendation for the routine use of anticholinesterase drugs in the treatment of patients with Alzheimer's disease

Evidence checklist

Level	RCTs (level II)
Quality	Good
Statistical precision	Small *P*-values
Size of effect	Small positive effect of treatment but clinical importance is doubtful
Relevance	Low (the duration of follow-up was too short in relation to the natural history of the disease and the outcomes measured were of doubtful relevance to patients and their carers).

Conclusion

Although the evidence was obtained from high quality RCTs, the other dimensions rated poorly and the recommendation cannot be supported.

2 APPLYING THE EVIDENCE

2.1 Introduction

The next step in formulating recommendations for clinical practice guidelines is to consider the applicability or generalisability of the evidence to patients or other population groups in relevant clinical settings. Before proceeding, the use of the terms *generalisability*, *external validity*, *extrapolation* and *applicability* should be clarified. They are often used interchangeably but actually have slightly different meanings. Definitions that will be used throughout this handbook are shown in Table 2.1.

Table 2.1 Definitions of terms relating to generalisability

Term	Definition
Generalisability (or external validity)	The extent to which a study's results provide a correct basis for generalisation beyond the setting of the study and the particular people studied. The application of the results to a group or population.
Extrapolation	The application of results to a wider population than that studied (ie to infer, predict, extend, or project beyond what was recorded, observed or experienced). For example, the results of a clinical trial in which patients aged 40–55 were studied, may be extrapolated to patients aged 55–65.
Applicability	The application of results to both individual patients and groups of patients. This addresses whether a particular treatment that showed an overall benefit in a study can be expected to convey the same benefit to an individual patient. In the clinical setting, applicability is preferred to the above terms as it includes the idea of particularising or individualising treatment and is closest to the general aim of clinical practice.

In the remainder of this section some broad issues will be described that should be considered first when assessing applicability. Methods described in the literature for assessing applicability and their flaws will be discussed briefly. The basis of these methods has been an attempt to apply the average treatment effect to people similar to those included in the studies. However it is argued that treatment decisions need to be individualised and the treatment should be targeted to patients in whom it is expected to do more good than harm. Finally, a process based on a risk –benefit approach for identifying individuals or groups in whom the intervention is likely to produce net benefit will be described.

Key points for applying the evidence

1. Broad issues regarding the applicability of the evidence should be considered. These include the reproducibility of the intervention in different settings and any biological factors that may alter the treatment's effectiveness (although these are likely to be rare).

2. Traditional methods for assessing applicability, based on subject selection for the trials, inclusion/exclusion criteria, and subgroup analyses, are flawed.

3. The aim is to identify, as the basis for treatment recommendations, for which individuals or homogeneous groups the treatment is more likely to do good than harm.

4. In order to do this, five questions need to be addressed:
   - (i) What are the beneficial and harmful effects of the intervention?
   - (ii) Are there variations in the relative treatment effect?
   - (iii) How does the treatment effect vary with baseline risk level?
   - (iv) What are the predicted absolute risk reductions for individuals?
   - (v) Do the benefits outweigh the harms?

2.2 General issues

At this stage of the process, the applicability of the evidence to other settings and to different patient groups should be considered in its broadest context. Issues such as whether the intervention can be reproduced in the settings of interest and the existence of biological factors that may make the treatment less effective are important.

2.2.1 Can the intervention be reproduced in the setting of interest?

This addresses the 'implementability' of the intervention and needs careful consideration. Generally, if the intervention is a drug then it is readily transferable to other settings. However, the drug has to be licensed for use and it must be affordable to the patients and health care system.

For a surgical procedure, skilled surgeons and support staff are required together with the necessary equipment. The implementation of complex interventions requires more careful consideration. For example, each component of a community health intervention program must be taken into account and be applicable in the new setting. This includes the availability of appropriately trained personnel, the sociocultural acceptability of the program content and so on (Gyorkos et al 1994). Also, the availability of the required infrastructure for any type of intervention should be considered. For example, for the prescription of warfarin to prevent stroke in patients with atrial fibrillation, the infrastructure for monitoring blood levels has to be in place and has to be accessible to patients. This may not be the case for patients living in rural and remote areas.

2.2.2 Can the results be applied in other settings and to other patients?

This refers to the transferability of the intervention. It is important to consider whether there are pathophysiological differences that may alter the biological mechanism through which the treatment has its effect. For example, could gender, age or ethnicity modify the effect? A well-known example where this is true is in the treatment of people with hypertension. The majority of trials of lipid-lowering drugs were carried out in middle-aged men. Many have debated whether the results apply to women or to older people.

The timing of the intervention should be considered. For example, the effect may be diminished if administration is delayed (eg thrombolytics after myocardial infarction) and so there may be no benefit beyond a certain time after the onset of symptoms or disease.

Variations in the nature of the disease may also change the effect of treatment. For example, geographical patterns of antibiotic-resistant bacteria and drug-resistant malaria may alter a treatment’s effectiveness. Also, the effect may be smaller in severe disease where the patient is very sick and the disease process has become irreversible.

These biological effect modifiers are probably not all that common, but they may be important in a few instances. However, the assumption should be made that, although they are unlikely, the burden of proof rests on demonstrating that they do modify the treatment effect rather than assuming that they may.

A situation where biological factors are important, however, is the generalisation of results about a treatment for one form of disease to another variant (see Box 2.1).

Box 2.1 Generalisation of diabetes results

The Diabetes Control and Complications Trial (DCCT) Research Group (1993) showed that intensive therapy and tight control for patients with insulin-dependent diabetes significantly reduced the occurrence of retinopathy, neuropathy, proteinuria and microalbuminuria.

However, there was also a three-fold increase in the incidence of severe hypoglycaemic reactions and an overall weight gain among those receiving intensive therapy.

As patients with noninsulin-dependent diabetes mellitus (NIDDM; or type II diabetes) tend to be middle-aged or elderly, they could not tolerate this frequency of hypoglycaemic episodes and their hyperinsulinaemia may be aggravated by weight gain (Marshall 1996b). Therefore intensive therapy with insulin for NIDDM could aggravate rather than ameliorate diabetic complications.

2.2.3 Can the results be applied to individuals?

The third aspect of applicability relates to applying the results to individuals. Of interest here is the effect of the underlying risk or prevalence of the condition of interest on the expected benefits and harms. A common argument against applying the results of clinical trials to clinical practice is that trials are typically carried out in tertiary care settings and so are not generalisable in community hospitals, general practice and other primary care settings (Sonis et al 1998).

However, as described in Sections 2.2.4 and 2.4.2, this can be addressed by characterising the individual (or patient groups) by their underlying risk and estimating the magnitude of the potential benefit, regardless of the health care delivery setting.

2.2.4 Using evidence in primary care

Patients in primary care settings are likely to have a lower baseline risk, and milder disease than those in tertiary care settings. There is therefore a belief that the biological response to treatment may be different and, in particular, that the treatment will be less effective.

This implies that the relative risk reduction varies with baseline (or underlying) risk of the event being prevented by the treatment. That is, the relative treatment effect is lower (or, in certain cases, it may be higher) in patients with lower baseline risk. This assumption needs to be checked as relative risk reduction may not change with baseline risk whereas absolute risk reduction (which is of more relevance to the individual) does change and is likely to be smaller for lower levels of baseline risk.

Therefore, it is important to consider the patient's baseline risk and a prediction of the absolute risk reduction. It may be the case that for a large proportion of patients in primary care, the expected benefits will not justify the potential harms and costs. However, there are likely to be some patients who will benefit from treatment. Therefore, the treatment should not be excluded from primary care entirely and individual patients should be assessed according to their baseline risk level.

The issues of relative versus absolute risk reductions and their relationships with baseline risk are described in greater detail in Section 2.4.

2.3 Applicability and guideline development

In Section 1, methods for assessing the evidence in terms of its strength (level, quality and statistical precision), size and relevance were described. The suggested output of this assessment is an evidence summary and checklist that collates the sources of the evidence (usually individual studies or systematic reviews), the main findings and a summary of the evidence on each dimension of the evidence (see Section 1.7).

At this point it should be reasonably clear whether the intervention has an overall beneficial effect. If it is beneficial, then the next thing to determine is, 'To whom should the intervention be offered?' It is very unlikely that the intervention, even if a statistically significant overall treatment effect has been established, will benefit everyone. This will be discussed in more detail later in this section.

After considering the more generic issues relating to the applicability of the results from individual studies or a systematic review of a number of studies, the next question that should be addressed is, 'How can the results be applied to individual patients?' The situation of particular interest here is choosing whether a preventive treatment is worthwhile. Preventive interventions can be divided into two groups —those with infrequent or minor adverse effects and those with frequent and/or serious adverse effects (Marshall 1996a).

Programs such as accident prevention, avoidance of high-risk behaviour, and healthy lifestyle changes fall into the first category. However, screening populations for disease, risk classification for selective preventive interventions and prophylactic drug treatment may be associated with more frequent and serious adverse effects. Because large numbers of people may be exposed to these interventions over long periods of time and for potentially small gains, it is important to establish in whom the benefits are likely to outweigh the harms.

In a recent systematic review of the literature, a number of methods for assessing the generalisability (or applicability) of results from RCTs and systematic reviews of RCTs were identified (O'Connell et al 1997). The main approaches described were based on a consideration of the criteria by which subjects were selected for inclusion, the source populations from which subjects were selected and how subjects included in the trials differed from those who would be candidates for the intervention but who were not included (or under-represented) in the trials.

Another approach is to examine the treatment effect in different subgroups defined by characteristics such as age, gender, ethnicity and severity of disease. The problems with this approach are that often the analysis is post hoc, the individual subgroups are too small to have sufficient power to detect a statistically significant treatment effect, and due to multiple hypothesis testing, a statistically significant effect may be detected in one subgroup by chance.

These approaches described in the literature (which are based on sampling theory and the representativeness of the subjects included in trials) are considered to be flawed. It is believed that by limiting the consideration of the applicability of evidence to these approaches, advances in the application of evidence to clinical and policy settings have been delayed.

A small number of papers in the literature described a more individualistic approach to applying the results of trials by considering the benefits and harms that might be expected from the intervention for individuals or groups with similar characteristics. This preferred approach for assessing the applicability of evidence is based on a risk–benefit analysis in which expected benefits and harms to an individual are assessed so that these can be weighed up when individual treatment decisions are being made (Lubsen and Tijssen 1989, Glasziou and Irwig 1995).

The rationale for this process is as follows. Even if there was a statistically significant and clinically important overall treatment effect in a trial, not all patients in the treatment group would have benefited from the intervention. Figure 2.1 represents the possible outcomes for a single patient. If we pretend that we can predict which patients will experience the event (eg stroke or death) that the treatment is trying to prevent, the following scenarios apply.

- If the patient was not likely to have the event, and did not, then the treatment has not bestowed any benefit and in fact may have caused harm through adverse effects.
- If the patient was deemed to have the event and did, then no benefit has been gained.
- If the patient would have had the event and, as a result of the treatment, did not, then the treatment has achieved some benefit.

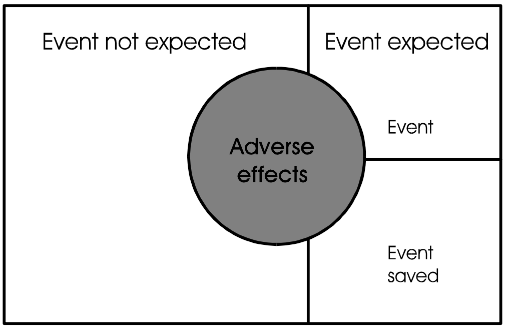


Figure 2.1 Possible outcomes for an individual in the treatment group of a randomised controlled trial

Therefore, the aim is to identify in which individuals the treatment will do more good than harm. A model proposed by Lubsen and Tijssen (1989) suggests that patient benefit (as measured by the prevention of an adverse event such as stroke or death) increases with (untreated) risk of the outcome, but harm or rates of adverse events caused by the treatment will remain relatively fixed (Figure 2.2). As the absolute risk reduction is related to the untreated (or baseline) risk of the event, high-risk patients are more likely to experience net benefit. Also, for some low-risk patients, the likely harms may outweigh the expected benefits.

This risk ~~–~~benefit approach was adapted to identify which patients with nonrheumatic atrial fibrillation may benefit from treatment with warfarin to prevent a stroke (Glasziou and Irwig 1995). A modified version of this approach is recommended for use and is described in Section 2.4.

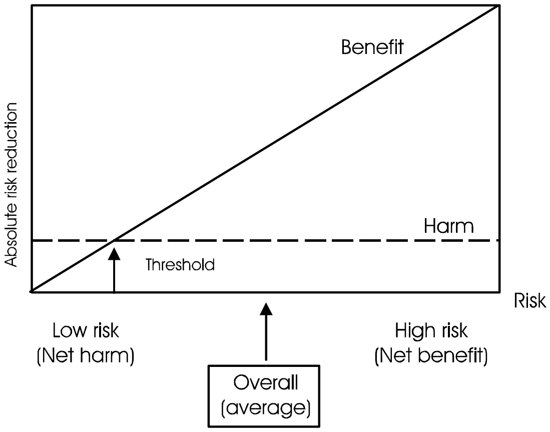


Figure 2.2 Risk–benefit approach to applying the evidence to individual patients (based on model of Lubsen and Tijssen 1989)

2.4 Applying evidence to individuals

The preferred approach for applying evidence to individuals is based on a risk – benefit model. In applying results of trials and systematic reviews in clinical practice, we need to consider whether there are predictors of individual response and risk, and how the risks and benefits balance. To do this, five questions need to be addressed (see Box 2.2) relating to transferability and application to individuals.

The first three questions shown in Box 2.2 relate directly to the overall effect of treatment —the expected outcomes (both benefits and harms) and the robustness of the estimate of (relative) treatment effect (ie is it constant across a range of potential effect modifiers and levels of baseline risk?). These address the issue of transferability of the average treatment effect.

The last two questions cover aspects of individualising or particularising the treatment decision through estimating the expected absolute risk reduction based on an individual's baseline risk and then taking into account the patient's preferences in the weighing up of benefits and harms. Each of these questions is discussed in detail below.

There will often be insufficient data to address each of the five questions in detail. However, it is still a helpful exercise to go through these steps and to make explicit what additional information is required to be able to apply the evidence about an intervention's effectiveness in an accurate manner. With improvements in trial design, and particularly in the analysis and reporting of trials, it will become possible to use this approach in an increasing number of situations.

2.4.1 Transferability

The first three steps involve the identification of all potential patient-relevant benefits and harms, identification of treatment effect modifiers and exploration of the relationship between the magnitude of the treatment effect and baseline (or untreated) risk. These all address the degree to which the estimates of treatment effect from a single study, or a systematic review of several studies, can be transferred to other settings and to other patients.

Box 2.2 Transferability of results and applicability to individuals

In applying results of trials and systematic reviews in clinical practice, five questions should be addressed.

TRANSFERABILITY

1. *What are the beneficial and harmful effects?*

All patient-relevant outcomes that are potentially influenced by the treatment, including, in particular, adverse effects, should be considered. It is helpful to begin by tabulating all that is known about possible positive and negative effects of the intervention.

2. *Are there variations in the relative treatment effect?*

Can the same (or average) treatment effect be applied to all subjects or does the effect vary according to varying characteristics of the patient, intervention, or disease? The evidence should come from testing whether the factor modifies the treatment effect (ie interaction), and not by testing within each individual 'subgroup'.

3. *How does the treatment effect vary with baseline risk level?*

Low-risk patients will usually gain less absolute benefit than high-risk patients. In order to estimate this, it is important to ascertain whether the relative effect varies with predicted event rate (or baseline risk).

APPLICATION TO INDIVIDUALS

4. *What are the predicted absolute risk reductions for individuals?*

To judge whether therapy is worthwhile, we need the absolute magnitude of benefit for an individual patient. This will vary with the patient's expected event rate (PEER): for low-risk patients absolute benefit may not outweigh the absolute harm. Thus to apply the results, the individual's PEER or severity based on established predictors is needed. Information on prognosis, external to the trials if possible, should be used.

5. *Do the benefits outweigh the harms?*

The absolute and net benefits of therapy, and the strength of the individual patient's preferences for these, need to be considered. The central issue is whether, for the individual patient, the predicted absolute benefit has greater value than the harm and cost of treatment.

Even if appropriate data are lacking, it will be helpful to think through these steps qualitatively.

Step 1 — What are the beneficial and harmful effects of the intervention?

The answer to the question 'To whom should the treatment be offered so that more good than harm is achieved?' requires careful consideration of all the potential benefits of the treatment as well as the harms (risks or side effects) that could be attributed to the treatment. The clinical guidelines development committee should therefore consider all patient-relevant endpoints that are potentially influenced by the treatment, including, in particular, adverse effects.

Of particular concern are patients who are at low risk of the main outcome that the intervention is supposed to prevent (eg death, stroke, myocardial infarction). Generally, in such patients the intervention can be expected to achieve only a small absolute risk reduction, so the harms are much more important as they are likely to outweigh the benefits.

The need for the measurement of all patient-relevant outcomes was discussed in Section 1. Typically a study or systematic review evaluating the effectiveness of the intervention focuses on a single primary endpoint and possibly a small number of secondary endpoints. Sometimes side effects are reported but often the number of participants with side effects is small because the trial was not designed to detect the effects of different treatments on these outcomes (particularly long-term effects).

It cannot be stressed enough that the clinical guidelines development committee should consider all patient-relevant endpoints that are potentially influenced by the intervention, including, in particular, adverse effects. For example, antiarrhythmic drugs have proarrhythmic effects; anticoagulants increase the risk of bleeding. Particularly for groups at low risk of the primary outcome such adverse effects may be crucial. It is helpful to begin the guideline development process by tabulating all that is known about possible positive and negative effects of the intervention, even if data for some outcomes are not available.

Examples of the potential benefits and harms associated with antihypertensive treatment in the elderly and screening of high-risk people for colorectal cancer are shown in Table 2.2 (Thijs et al 1994, Eddy 1990b).

Table 2.2 Potential benefits and harms associated with preventive treatments for hypertension and colorectal cancer

Benefits	Harms	
Treatment of hypertension in the elderly[a]		
Reduction in mortality due to all causes (noncardiovascular, cardiovascular, coronary, cerebrovascular). Reduction in nonfatal stroke.	Gout Muscle cramps Dizziness Dyspnoea Dry mouth	Skin disorders Nausea Rynaud's phenomenon Headaches Diarrhoea
Screening high-risk individuals for colorectal cancer[b]		
Reduced risk of invasive colorectal cancer. Reduced risk of dying from colorectal cancer.	False positive result leading to a clinical workup Perforation of the colon Discomfort Inconvenience Anxiety	

[a] Thijs et al 1994
[b] Eddy 1990b

Step 2 — Are there variations in the relative treatment effect?

Chance variation between subgroups is inevitable; hence without prior justification and strong evidence, we should assume there is no variation. The evidence should come from testing whether the factor modifies the treatment effect (ie interaction), and not by testing within each individual 'subgroup'. Ideally this is done from individual data *within trials* (not between trials), otherwise confounding by variation in trial design may occur.

To better understand how the intervention works, and how to apply the results of evaluations, it is useful to determine whether treatment effect is constant or whether it varies according to various characteristics of the patients (eg age, gender, biochemical markers); the intervention (eg the timing, compliance, or intensity of the intervention); the disease (eg hormone receptor status); and the measure of effect used (relative risk versus risk difference).

A distinction should be made between 'effect modification', 'heterogeneity' and 'interaction', which tend to be used interchangeably, but which have different meanings. The definitions used throughout this handbook are given in Table 2.3.

Table 2.3 Variations in relative treatment effect: definitions of heterogeneity, interaction and effect modification

Cause of variation	Definition
Heterogeneity	Differences in estimates of treatment effect between studies contributing to a meta-analysis. Significant heterogeneity suggests that the trials are not estimating a single common treatment effect. The variation in treatment effects may be due to differences in the patients included in the trials, the setting, the way the intervention was administered, the way outcomes were defined and measured and so on.
Interaction and effect modification	The relationship between a single variable (or covariate) and the treatment effect. Significant interaction between the treatment and a variable (such as patient's age, patient's gender or ethnicity) indicates that the treatment effect varies across levels of this variable. If this is the case, then the variable is called an 'effect modifier' as it modifies the effect of treatment. This modification of effect may contribute to any heterogeneity detected.
Qualitative interaction	Where the treatment effect differs in direction across levels of the covariate. For example, the treatment may reduce risk of the outcome in men but increase risk in women.
Quantitative interaction	Occurs when the magnitude of the treatment effect varies but is in the same direction. For example, the relative risk reduction is 50% in men and 30% in women. It has been argued that quantitative interactions are much more likely to occur than qualitative interactions because humans are not all that different biologically. Therefore a significant qualitative interaction may be an extreme result.

When there are several studies evaluating the intervention, the treatment effect can be explored further by firstly testing for heterogeneity using methods described in the accompanying handbook on systematic literature review (NHMRC 2000a). If the treatment effect appears to be heterogeneous (or nonconstant) then the sources of heterogeneity (or important effect modifiers) should be examined by testing for interactions between each of the variables and treatment. Regardless of the result of the test of heterogeneity, interactions should be explored, since the test of heterogeneity typically has low power and sometimes heterogeneity may not be apparent even in the presence of effect modification (because the studies were homogeneous on the effect modifier). Advice from a biostatistician should be sought on how best to do this.

Potential effect modifiers vary according to the treatment and disease under investigation. For example, a number of biological/disease factors and patient features may be effect modifiers for treatment of breast cancer, including age/menopausal status, nodal status, receptor status, tumour size, grade, ploidy and type, vascular invasion, labelling index, S-phase fraction, body mass and family history (Gelber and Goldhirsch 1987). Heterogeneity may occur by chance (but this can be checked using a statistical test) or by choosing the wrong measure of treatment effect (relative risk versus risk difference). If these causes have been dismissed then possible sources of heterogeneity can be grouped into four areas, which are shown with some examples in Table 2.4. An example of the exploration of effect modification is described in Box 2.3

Table 2.4 Potential sources of heterogeneity in treatment effect (adapted from Guyatt et al 1995)

Potential effect modifier	Examples
Trial design	setting, subjects, co-administered therapy, length of follow-up, outcomes measured and method used, completeness of follow-up, study quality
Characteristics of:	
– patient	age, gender, race, comorbidity, biochemical markers, genetic markers
– disease	method and accuracy of diagnosis, severity, stage, responsiveness
– intervention	form, mode of administration, dose, intensity, timing, duration, compliance

Step 3 — How does the treatment effect vary with baseline risk level?

A very important factor to consider when looking at whether the magnitude of the benefit varies across different groups is the level of (untreated) risk. Generally low-risk groups will have less to gain from an intervention than high-risk groups, and may not therefore gain sufficient benefit to outweigh harms from treatment (adverse effects, costs, etc).

Box 2.3 Are there variations in the relative treatment effect?

Study: Effect modification by gender was examined in an individual patient data meta-analysis of the effect of antihypertensive drug treatment on cardiovascular outcomes (Gueyffier et al 1997).

Subgroup analysis
Subgroup analyses suggested that while statistically significant reductions in risk due to treatment were found for all seven outcomes in men, the odds ratios were statistically significant in women for only three of the outcomes.

Test for interaction
No significant interaction was found between gender and (relative) treatment effect: the odds ratio comparing the treatment and control groups did not differ significantly between men and women for any of the seven outcomes analysed.

Conclusion
The subgroup analyses suggested (incorrectly) that treatment was less effective in women. The appropriate analysis was the test for interaction, which suggested that gender is not an effect modifier and that antihypertensive treatment was equally effective in men and women.

The benefit for an individual (as measured by the absolute risk reduction or NNT) is likely to depend strongly on the baseline level of risk. For example, a 50% relative risk reduction could mean a reduction in risk from 10% to 5% (a 5% absolute risk reduction or NNT of 20) or from 1% to 0.5% (a 0.5% absolute risk reduction or NNT of 200). Which patients obtain a net benefit depends on the harmful effects of the intervention, such as adverse drug reactions, side effects, the burden of compliance and monitoring, etc. The greater the potential harm, the greater the benefit required to make it worthwhile. This trade-off and how it relates to baseline risk is illustrated in Figure 2.2 which shows increasing benefit but constant harm across different levels of baseline risk.

This extrapolation from the estimated relative risk to the expected absolute risk reduction for different risk groups assumes that the relative risk reduction remains constant —*an assumption that needs to be checked*. For many interventions the relative risk appears to be reasonably constant. An analysis of individual study results included in 112 meta-analyses found a statistically significant relationship between the measure of treatment effect and baseline risk in 15 (13%) meta-analyses with relative risk as the measure of treatment effect, in 16 (14%) with the odds ratio and 35 (31%) with the risk difference (Schmid et al 1998).

Thus, the relative risk was constant for 87% of the systematic reviews. However, there are clearly cases that deviate considerably from this assumption.

For example, Rothwell (1995) showed that for patients with carotid artery stenosis treated with carotid endarterectomy, the relative risk was very different for different risk groups. In the high-risk group there was a high relative risk reduction, but in the low-risk group there was no apparent benefit. By contrast, he showed that aspirin had a similar relative benefit across both groups. A second example of a nonconstant relative risk reduction is class I anti-arrhythmic drugs. Boissel et al (1993) found that for trials of treatment of patients after myocardial infarction with class I anti-arrhythmics, there was a large variation in the degree of risk in these trials because of different inclusion criteria. There appeared to be a beneficial effect in the very high-risk group, but in lower-risk groups, the drugs appeared to cause net harm. This result was confirmed by CAST (1989), which showed a doubling of mortality from the anti-arrhythmic drug flecainide.

Thus, because most patient groups to which trial results are applied do not have a risk identical to the average seen in the trial, a crucial analysis in assessing the applicability of trial results is a systematic study of how relative benefit varies with baseline risk. This may be done either within trials, by looking at different prognostic groups, or between trials, by making use of the variation in average baseline risk due to the different selection processes of trials. An example of nonconstant relative risk is described in Box 2.4 (Example 1).

A number of methodological problems can produce misleading results in such analyses. For example, Brand and Kragt (1992) suggested plotting the relative risk against the risk in the control group (as a surrogate for the underlying baseline or untreated risk). In their example —tocolysis for preterm birth — relative risk reduction appeared to decrease with decreasing control group risk (ie a nonconstant relative risk). While such an analysis is potentially very informative, there are methodological pitfalls, which have been discussed by a number of authors (McIntosh 1996, Sharp et al 1996, Walter 1997, Thompson et al 1997). Advice from a statistician may be required for such an analysis.

Also, when the analysis is undertaken at the study level (eg the average risk in the control group is compared with the average treatment effect for each study), potential confounding by other factors that vary across the studies may produce a spurious association (analogous to the ecological fallacy).

The other assumption is that the harms are constant across levels of baseline risk for the event being prevented. In many cases, this is likely to be true as the harms are associated with exposure to the intervention itself. However, this assumption should be checked. An example is shown in Box 2.4 (Example 2).

Box 2.4 How does the treatment effect vary with baseline risk level?

Example 1 — nonconstant relative risk

Study

A systematic review of 39 RCTs of angiotensin converting enzyme (ACE) inhibitors in the treatment of congestive heart failure (North of England Evidence-based Guideline Development Project) (NEEGDP 1997). The studies were stratified according to the annual mortality rate:

- low-risk group — mortality of up to 15% (28 studies); and
- high-risk group — mortality greater than 15% (11 studies).

Result

Relative risk for mortality in the group treated with ACE inhibitors compared to those receiving placebo:

- 0.88 (95% CI 0.80 to 0.97) for the low-risk group; and
- 0.64 (95% CI 0.51 to 0.81) for the high-risk group.

Conclusion

Both the absolute and relative risk reductions due to use of ACE inhibitors in patients with impaired left ventricular dysfunction and symptoms of heart failure increased with mortality risk.

Example 2 — constant relative risk

Study

A systematic review of 16 trials in which patients were randomised to aspirin or a control group for at least one month and compared for cardiovascular events and haemorrhagic stroke (He et al 1998, Boissel 1998).

Result

The relative risk reduction of ischaemic cardiovascular events (benefit) was constant over a wide range of baseline risks. Further, the absolute risk increase in haemorrhagic stroke (harm) was 12 events per 10,000 people treated and this risk did not differ by patient or study design characteristics. Therefore, the excess absolute risk of haemorrhagic stroke is constant and independent of the patient's absolute risk of cardiovascular events.

Conclusion

The absolute benefit is proportional to the patient's baseline risk for developing an ischaemic cardiovascular event against an underlying constant risk of harm. Therefore high-risk patients will derive greater net benefit from aspirin.

2.4.2 Application to individuals

The next two issues to be considered relate directly to applying the results to an individual (or a group of similar individuals). Working through these steps and including the results in the clinical practice guidelines will greatly assist clinicians and patients make informed treatment decisions.

Step 4 — What are the predicted absolute risk reductions for individuals?

While the relative risk is useful for assessing the biological strength of response to the treatment, to judge whether therapy is worthwhile for an individual, the absolute magnitude of benefit should be estimated. This might be expressed as the absolute risk reduction, or as the number needed to treat (NNT). However it is expressed, it varies with the patient's baseline risk (which is sometimes referred to as the patient's expected event rate, PEER): for low-risk patients, absolute benefit may not outweigh the absolute harm. If the relative risk reduction is constant across levels of baseline risk and there is no effect modification, then the average relative risk reduction can be applied to a particular patient. If this is not the case, then the relationships need to be described so that the appropriate relative risk can be applied to a particular patient.

To estimate the absolute risk reduction for an individual, the expected relative risk reduction can be applied to the PEER. This requires an estimate of PEER that can be obtained from a previously developed (and preferably validated) prognostic model linking values of various (baseline/prerandomisation) characteristics of the patient to the probability of the disease of interest.

The process of estimating the PEER and the corresponding absolute risk reduction for an individual patient is described in Box 2.5.

There are many examples of prognostic models in the literature (some are described in Table 2.5). They tend to be based on cohort (prospective) studies in which a sample of patients are recruited, various baseline characteristics are recorded and then the patients are followed over time to see whether they develop the outcome of interest.

Statistical modelling (such as logistic regression or proportional hazards models) is then used to examine which combination of variables best predicts which patients will experience the outcome of interest.

Box 2.5 What are the predicted absolute risk reductions for individuals?

In considering whether to treat two patients with aspirin to prevent a further cardiovascular event, Boissel (1998) estimated their one-year mortality risk using a previously established risk score.

Patient 1: 45-year-old man with normal blood cholesterol, glucose levels, blood pressure and weight and who is a non-smoker.
Presented with an uncomplicated inferior myocardial infarction.
One-year mortality risk = 2%.

Patient 2: 65-year-old woman with hypertension for 30 years, diabetes, a previous myocardial infarction, smokes two packs of cigarettes a day and reports breathlessness with exercise.
Presented with her second anterior myocardial infarction together with signs of heart failure.
One-year mortality risk = 30%.

Applying a (constant) relative risk reduction of 15% in all-cause mortality, the two patients can expect absolute risk reductions of 0.3% and 4.5%, respectively. These can be weighed against an excess risk of cerebral bleeding with aspirin of 0.12% for both patients.

Step 5 — Do the benefits outweigh the harms?

The absolute and net benefits of therapy, and the strength of the individual patient's preferences for these, need to be considered. If the treatment has multiple effects, for example, adverse as well as beneficial effects, then the assessment of the absolute benefit needs to incorporate these disparate outcomes. If step 4 is done well, the trade-offs will often be clear; however, methods developed in decision analysis may be a useful supplement, for example, quality-adjusted life-years (QALY) might provide a summary measure when there are trade-offs between quality and quantity of life. The central issue is whether, for the individual patient, the predicted absolute benefit has greater value than the harm and cost of treatment. For example, when does the reduction in strokes outweigh the risk of bleeding from anticoagulation (see Box 2.6); or when does the benefit of surgery outweigh its risk?

Table 2.5 Examples of prognostic models from the literature

Author/ year	Outcome	Predictor variables	Sample	Validated	Presentation
Schuchter et al (1996)	10-year survival after definitive surgical therapy for primary cutaneous melanoma	Tumour thickness, primary melanoma, patient's age, sex	Patients at university medical centre	Yes (on 142 patients)	Table of probabilities of 10-year survival
Simoons and Arnold (1993)	1-year mortality after myocardial infarction	Advanced age, history of previous infarction, anterior location of current infarction, heart failure, QRS duration >120 ms, total ST segment deviation[a]	Data from 3 trials of thrombolytics in advanced myocardial infarction	No	Table of probability of death
Simoons et al (1993)	Probability of intracranial haemorrhage after thrombolytic therapy	Age >65, weight <70 kg, hypertension drug regimens with alteplase	150 patients with documented intracranial haemorrhage, 294 matched controls	No	Table of probabilities of intracranial haemorrhage assuming 50% or 75% overall incidence of intracranial haemorrhage
NZCSC (1995)[b]	5-year risk of a cardiovascular event	Gender, age, blood pressure, diabetes, smoking, total:HDL cholesterol ratio	Framingham data	No	Coloured nomograms for men and women separately
SPAFI (1994)[c]	Rates of thrombo-embolic stroke in patients with atrial fibrillation	Hypertension, recent congestive cardiac failure, previous thromboembolism, left ventricular dysfunction atrial size	Control group of SPAF trial (211 patients)	No	Annual risk of stroke

[a] Measured on an electrocardiogram trace

[b] Guidelines for the management of mildly raised blood pressure in New Zealand (NZCSC 1995)

[c] Stroke Prevention in Atrial Fibrillation Investigators

Box 2.6 Trade-off between benefit and harm

Clinical problem

Prescribing warfarin for patients with nonvalvular atrial fibrillation for the prevention of stroke (Stroke Prevention in Atrial Fibrillation Trial; SPAF Investigators 1994).

Benefits

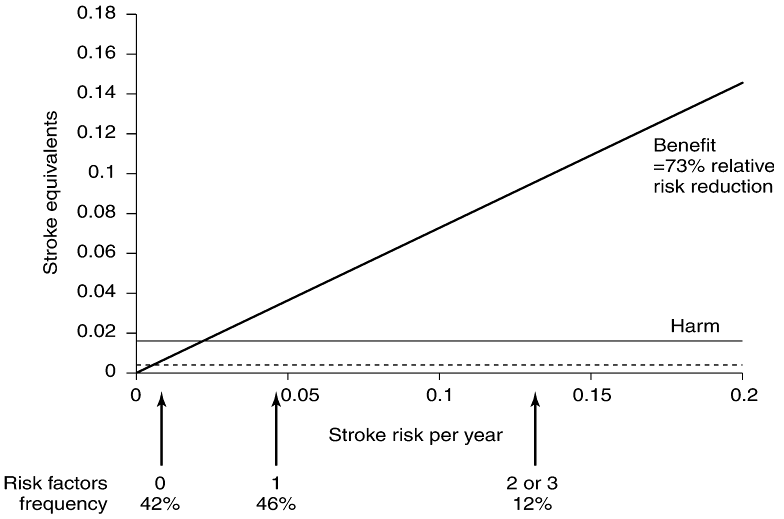
73% relative risk reduction for stroke.

Harms

Absolute increase in risk of (fatal) intracranial haemorrhage of 1–3%.

Trade-off

The figure illustrates the trade-off between benefit and harm for patients at different levels of baseline risk, as indicated by the number of risk factors present.



Previous work (Glasziou and Irwig 1995) suggests that patients consider a fatal intracranial haemorrhage to be equivalent to four strokes (the upper horizontal line in the figure).

Therefore the harms are likely to outweigh the benefits in patients with none of the three risk factors (hypertension, recent congestive cardiac failure, previous thromboembolism), which included 42% of the patients in the Stroke Prevention in Atrial Fibrillation Trial (SPAF Investigators 1994). It is only for patients with one or more risk factors that the benefit may outweigh the potential harmful effect of an intracranial haemorrhage.

The final consideration in formulating recommendations about to whom the intervention should be offered is how the potential benefits and harms compare. In addition to formulating treatment recommendations, all information should be presented on the benefits and harms by levels of baseline risk so that individuals (patients) together with their health care providers can, if they choose, make a decision about the intervention based on their preferences.

Often, there will be multiple effects associated with the intervention —both beneficial and harmful. These need to be weighed up and trade-offs made in formulating recommendations and in making individual treatment decisions.

If no adverse events have been identified, then it may seem that only benefits need to be considered. However, serious adverse events may be rare and hence the concern is how reliably they have been excluded. This 'proof of safety' is usually more difficult, and rare adverse effects (say of drugs) may only emerge during post-marketing monitoring (see Box 2.7). Several authors have discussed the statistical difficulties of excluding such serious rare events. For example, if 300 patients have been treated in controlled trials and no serious adverse effects detected, the confidence interval for the rate of such events is from 0% to 1% —that is, we can be 95% certain that the risk of an adverse effect is less than 1% (Hanley and Lippman-Hand 1983).

Box 2.7 Appearance of adverse effects during post-marketing monitoring of a drug

Mibefradil, a T-type and L-type calcium channel blocker was released in the United States in 1997 for the management of hypertension and chronic stable angina (Mullins et al 1998).

It was listed on the Pharmaceutical Benefits Scheme in Australia in May 1998.

In June 1998, mibefradil was withdrawn from the market worldwide due to potentially serious interactions with many commonly prescribed drugs, including lipid-lowering statins and beta-blockers, which would be commonly coprescribed to patients with cardiovascular disease (Krum and McNeil 1998).

These adverse effects were not detected in the studies of its efficacy and safety, due to insufficient sample sizes. These effects did not become apparent until the drug was released into large markets through ad hoc post-marketing surveillance.

The trade-offs between different outcomes can involve either the quality of the outcome, such as different dimensions of quality of life, or they may involve the timing of occurrence of the outcome. A particularly frequent and important trade-off between outcomes is that between quality and quantity of life. For example, many surgical procedures, such as hip replacement or cholecystectomy, involve a mortality risk that may be acceptable because of the reduction in symptoms or improvement of function. Appropriate reporting and analysis of clinical trials is needed to make the trade-offs explicit, and to allow an informed decision by the health care provider and patient. The different types of trade-off that may be required for the intervention of interest should be made explicit.

However, if step 4 — estimating an individual's potential benefits and harms — is done well, and presented clearly, the trade-offs will often be clear. The central issue is whether, for the individual patient, the predicted absolute benefit has greater value than the harm and cost of treatment.

3 PRESENTING THE EVIDENCE ON BENEFITS AND HARMS

3.1 Introduction

In addition to making treatment recommendations based on the review of the evidence, the data should be presented in such a way that if clinicians and patients choose to consider the evidence, and to weigh up the benefits and harms in light of their personal preferences, sufficient information is provided to do so. The effect that different presentation formats, in particular relative risk versus risk difference or number needed to treat (NNT), can have on the perceived worth of an intervention should be considered when preparing these summaries.

In addition to the evidence summary and checklist summarising the evidence and its strength (level, quality and statistical precision), size of effect and relevance, a quantitative summary of the expected benefits and harms (in both relative and absolute terms) should be prepared. Following recommendations from Eddy (1990b) and Liddle et al (1996), it is strongly advised that a table of the evidence be presented that represents a 'balance sheet' of the benefits and harms (see Table 3.1). This section describes how this is done. The effects of 'information framing', or how the data are presented, are also described briefly.

Key points for presenting evidence on benefits and harms

1. The way in which data about the size of the treatment effects are framed can influence perceptions about the treatment's worth.

2. Summary data should be presented on the treatment effect in both relative and absolute terms.

3. Estimates of absolute risk reduction and NNT should be provided for a range of values for baseline risk.

4. A table or nomogram should be included to assist clinicians and patients estimate the patient's baseline risk.

5. A balance sheet summarising data on both benefits and harms will aid an accurate understanding of the consequences of different treatment options.

3.2 Information framing

There is some evidence to suggest that the way in which the treatment effect is expressed (or 'framed'), influences the perceived effectiveness of the intervention. In particular, it has been suggested that the intention to prescribe a treatment is higher when its effect is expressed in relative terms (ie as a relative risk) compared to when it is expressed as the absolute risk reduction or NNT (McGettigan et al 1999).

For example, a 50% relative risk reduction sounds more impressive than a 0.5% (or 5 in 1000) reduction in risk (from 1% to 0.5%) or an NNT of 200.

Summaries of data about the treatment effect should, therefore, be presented in both relative and absolute terms as they convey different information and provide a better overall understanding of the potential treatment benefits. The relative risk is useful for assessing the biological strength of the response to the treatment. However, it has little meaning for the individual as it does not take into account their baseline (or starting) risk.

On the other hand, the absolute risk reduction or NNT, since it is dependent on baseline risk, is not meaningful on its own and it requires additional information about the baseline risk for which it is has been calculated. The absolute risk reduction and NNT also vary with duration of treatment and follow-up, so a time period should be specified when quoting these results. For example, a patient with congestive heart failure admitted to hospital has a predicted probability of dying in less than 28 days of 16% and of dying within one year of 73%. The use of ACE inhibitors reduces this risk by 17% (in relative terms) (NEEGDP 1997) which corresponds to absolute risk reductions of 2.7% (NNT of 37) at 28 days and 12% (NNT of 9) at 12 months. This dependence on baseline risk and time means that studies are more likely to exhibit heterogeneity in treatment effect when the risk difference is used.

Therefore, it is recommended that the relative risk is used to examine constancy of treatment effect across effect modifiers and baseline risk and then the relative risk should be converted to absolute risk reductions and NNTs corresponding to different levels of baseline risk.

3.3 Presenting a balance sheet of benefits and harms

3.3.1 What is a balance sheet?

The balance sheet is a mechanism through which the information obtained about the effect of the intervention based on the five questions posed in

Section 2 can be summarised and presented. The purpose of the balance sheet is:

- to provide justification for the treatment recommendations made in the clinical practice guidelines; and

- to aid clinical decision making when the clinician and patient opt to consider the evidence and weigh up the benefits and harms, taking into account personal preferences.

3.3.2 What should be included in the balance sheet?

The balance sheet should present data on both benefits and harms for all the relevant outcomes.

An example of a balance sheet is shown in Table 3.1, which summarises the relevant data on the effect of screening a high-risk person with annual faecal occult blood tests and 60-cm flexible sigmoidoscopies every three years from age 50 to age 75 years (Eddy 1990b).

When this balance sheet was constructed, there was no direct evidence from RCTs that screening decreased mortality from colorectal cancer. Instead, the data shown in the balance sheet were derived from information on incidence rates, the natural history of colorectal cancer and the effectiveness of the different screening tests. This emphasises the importance of the accompanying evidence summary and checklist which acts as a reminder to users of the guideline that this is not based on high level evidence. Recent work suggests that screening for colorectal cancer using a biennial faecal occult blood test is associated with an overall 16% relative risk reduction for dying with colorectal cancer (23% in those actually screened) (Towler et al 1998). Therefore, for a 50-year-old man with a baseline risk of dying of colorectal cancer of 5.3%, an absolute reduction in risk of –1.2% would be expected. This is less than Eddy's estimate but annual faecal occult blood tests and three-yearly sigmoidoscopy may provide added benefit.

While Eddy (1990b) included the costs of screening and treatment in his balance sheet, these have been omitted here as they reflect the costs of the health care system in the United States. It should also be noted that this balance sheet presents the benefits and harms for an individual at a specific level of baseline risk. Therefore, multiple balance sheets (one for each major risk category) would be required, or the results corresponding to different baseline risks could be included in a single table.

Table 3.1 Example of benefit–harms balance sheet for a colorectal cancer screening strategy (high-risk 50-year-old man with screening to age 75 years)

Outcome	No screening	FOBT + scope[a]	Difference due to screening	Range of uncertainty
Benefits				
Probability of getting colorectal cancer	10.3% (103/1000)	7.3% (73/1000)	–3.0% (30/1000)	0–6%
Probability of dying of colorectal cancer	5.3% (53/1000)	2.9% (29/1000)	–2.4% (24/1000)	0–5%
Probability of having undetected cancer[b]	0.1% (1/10 000)	0.03% (3/10 000)	–0.07% (7/10 000)	0.05–0.09%
Harms				
Probability of false positive FOBT	0%	40% (400/1000)	40% (400/1000)	
Probability of perforation[c]	0%	0.3% (3/1000)	0.3% (3/1000)	
Inconvenience/anxiety/ discomfort[d]	–	–	–	–
Number of FOBTs	0	26 procedures	+26 procedures	
Number of sigmoidoscopies	0	9 procedures	+9 procedures	

[a] Faecal occult blood test (FOBT) every year; 60-cm flexible sigmoidoscopy (scope) every three years
[b] Probability that the patient has a cancer that will develop signs or symptoms in the coming year
[c] Probability of perforation due to work up after false positive test
[d] These factors should be taken into consideration but are difficult to quantify
Source: adapted from Eddy 1990b

For completeness, a nomogram or table of the estimated baseline risks corresponding to different risk factor profiles calculated from a suitable prognostic equation should be included. The coloured nomograms from the New Zealand guidelines for the management of hypertension are a good example (NZCSC 1995[1]). This will aid clinician and patient to identify the relevant level of baseline risk and hence the magnitude of the expected benefits.

[1] Updated versions are available on the Internet at: http://cebm.jr2.ox.ac.uk/docs/prognosis.html

APPENDIX A MEMBERSHIP OF PRODUCTION TEAM FOR HANDBOOK

NHMRC Assessment Panel

Professor Paul O'Brien (Chair)	Department of Surgery, Monash Medical School Member of the NHMRC Health Advisory Committee (HAC)
Professor Chris Silagy	Monash Institute of Public Health and Health Services Research Member of HAC
Professor John McCallum	Faculty of Health, University of Western Sydney Member of HAC

Consultant authors*

Dr Dianne O'Connell	Discipline of Clinical Pharmacology, University of Newcastle, New South Wales
Associate Professor Paul Glasziou	Department of Social and Preventive Medicine, University of Queensland
Dr Suzanne Hill	Discipline of Clinical Pharmacology, University of Newcastle, New South Wales

Technical writer / editor

Dr Janet Salisbury	Biotext, Canberra

Secretariat

Ms Roz Lucas, Ms Janine Keough, Ms Monica Johns	Health Advisory Unit, Office of NHMRC

* The authors would also like to acknowledge the helpful comments and input of Professor David Henry and Professor Bruce Armstrong during the development of this handbook.

APPENDIX B

PROCESS REPORT

During the 1997–99 NHMRC triennium the Health Advisory Committee focused its work on the areas of coordination and support rather than on collating and reviewing scientific evidence. However, the committee recognised that a key part of its coordination and support function was to provide a methodology on how to develop evidence-based guidelines.

The NHMRC publication *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999), which had been produced by the Health Advisory Committee as a resource for people wishing to develop clinical practice guidelines to a standard acceptable to the NHMRC, was revised during 1998. Early in the revision process, the committee realised that there was a need for a number of complementary handbooks to expand on the principles outlined in the document. This complementary series would cover other aspects of the identification, collation and application of scientific evidence. It was envisaged that these handbooks would be of invaluable assistance to agencies wishing to develop clinical practice guidelines of a high standard either independently, or on behalf of the NHMRC.

It was agreed that there would initially be five handbooks in the series:

- how to review the evidence;
- how to use the evidence;
- how to put the evidence into practice;
- how to present the evidence for consumers; and
- how to compare the costs and benefits.

They would be published individually to allow flexibility in their production and revision, as well as to allow any later additions to the series.

Recognising the need for a transparent and competitive process for contracting the services of an expert(s), tenders were sought for the preparation of each handbook. A selection committee was then appointed by the Health Advisory Committee to consider the tenders.

Once the successful tenderers had been contracted to prepare the handbooks, an assessment panel, composed of Health Advisory Committee members, was formed to manage the progress of each project (see Appendix A).

When first drafts of each handbook were received, they were distributed to a small number of experts in that particular field for peer review. The documents were subsequently revised in the light of these comments. A technical writer was employed to ensure consistency in content and style within and between the handbooks.

The finalised documents were referred, in turn, to the Health Advisory Committee for approval before being forwarded to the NHMRC for endorsement.

GLOSSARY

Absolute risk reduction

The effect of a treatment can be expressed as the difference between relevant outcomes in the treatment and control groups by subtracting one rate (given by the proportion who experienced the event of interest) from the other. The reciprocal is the number needed to treat (NNT).

Accuracy (*see also* validity)

The degree to which a measurement represents the true value of the variable which is being measured.

Adverse event

A nonbeneficial outcome measured in a study of an intervention that may or may not have been caused by the intervention.

Adverse reaction

Any undesirable or unwanted consequence of a preventive, diagnostic or therapeutic procedure.

Allocation (or assignment to groups in a study)

The way that subjects are assigned to the different groups in a study (eg drug treatment/placebo; usual treatment/no treatment). This may be by a random method (*see* randomised controlled trial) or a nonrandom method (*see* pseudorandomised controlled study).

Applicability (*see also* external validity, generalisability)

Encompasses the application of results to both individual patients and groups of patients. This is the preferred term as it includes the idea of particularising or individualising treatment and is closest to the general aim of clinical practice. It addresses whether a particular treatment that showed an overall benefit in a study can be expected to convey the same benefit to an individual patient.

Baseline risk

An estimate of an individual patient's (untreated) risk of an outcome.

Before-and-after study (*see* pretest/post-test study)

Bias

Bias is a systematic deviation of a measurement from the ‘true’ value leading to either an over- or underestimation of the treatment effect. Bias can originate from many different sources, such as allocation of patients, measurement, interpretation, publication and review of data.

Blinding

Blinding or masking is the process used in epidemiological studies and clinical trials in which the observers and the subjects have no knowledge as to which treatment groups subjects are assigned. It is undertaken in order to minimise bias occurring in patient response and outcome measurement. In single-blind studies only the subjects are blind to their allocations, whilst in double-blind studies both observers and subjects are ignorant of the treatment allocations.

Case-control study

Patients with a certain outcome or disease and an appropriate group of controls without the outcome or disease are selected (usually with careful consideration of appropriate choice of controls, matching, etc) and then information is obtained on whether the subjects have been exposed to the factor under investigation.

Case series

The intervention has been used in a series of patients (may or may not be consecutive series) and the results reported. There is no separate control group for comparison.

Causality

The relating of causes to the effects they produce. The Bradford–Hill criteria for causal association are: consistency, strength, specificity, dose – response relationship, temporal relationship (exposure always precedes the outcome – it is the only essential criterion), biological plausibility, coherence and experiment.

Clinical outcome

An outcome for a study that is defined on the basis of the clinical outcome being studied (eg fracture in osteoporosis, peptic ulcer healing and relapse rates).

Clinically important effect (*see also* statistically significant effect)

An outcome that improves the clinical outlook for the patient. The recommendations made in clinical practice guidelines should be both highly statistically significant *and* clinically important (so that the 95% CI includes clinically important effects).

Cochrane Collaboration

The Cochrane Collaboration is an international network that aims to prepare, maintain and disseminate high quality systematic reviews based on RCTs and when RCTs are not available, the best available evidence from other sources. It promotes the use of explicit methods to minimise bias, and rigorous peer review.

Cohort study

Data are obtained from groups who have been exposed, or not exposed, to the new technology or factor of interest (eg from databases). Careful consideration is usually given to patient selection, choice of outcomes, appropriate controls, matching, etc. However, data on outcomes may be limited.

Comparative study

A study including a comparison or control group.

Concurrent controls

Controls receive the alternative intervention and undergo assessment concurrently with the group receiving the new technology/intervention. Allocation to the intervention or control is not random.

Confidence interval (CI)

An interval within which the population parameter (the ‘true’ value) is expected to lie with a given degree of certainty (eg 95%).

Confounding

The measure of a treatment effect is distorted because of differences in variables between the treatment and control groups that are also related to the outcome. For example, if the treatment (or new intervention) is trialed in younger patients then it may appear to be more effective than the comparator, not because it is better, but because the younger patients had better outcomes.

Cross-sectional study

A study that examines the relationship between diseases (or other health-related characteristics) and other variables of interest as they exist in a defined population at one particular time (ie exposure and outcomes are both measured at the same time).

Cumulative meta-analysis (*see also* meta-analysis)

In a systematic review, the results of the relevant studies are ordered by some characteristic and sequential pooling of the trials is undertaken in increasing or decreasing order.

Double-blind study (*see* blinding)

Ecological fallacy

The bias that may occur because an association observed between variables on an aggregate (eg study or country) level does not necessarily represent the association that exists at an individual (subject) level.

Effectiveness

The extent to which an intervention produces favourable outcomes under usual or everyday conditions.

Effect modification, effect modifier (*see also* interaction)

The relationship between a single variable (or covariate) and the treatment effect. Significant interaction between the treatment and such a variable indicates that the treatment effect varies across levels of this variable.

Efficacy

The extent to which an intervention produces favourable outcomes under ideally controlled conditions such as in a randomised controlled trial.

Evidence

Data about the effectiveness of a new treatment or intervention derived from studies comparing it with an appropriate alternative. Preferably the evidence is derived from a good quality randomised controlled trial, but it may not be.

Evidence-based medicine/health care

The process of finding relevant information in the medical literature to address a specific clinical problem. Patient care based on evidence derived from the best available studies.

External validity (*see also* generalisability, applicability)

Also called generalisability, is the degree to which the results of a study can be applied to situations other than those under consideration by the study, for example, for routine clinical practice.

Extrapolation

Refers to the application of results to a wider population and means to infer, predict, extend, or project the results beyond that which was recorded, observed or experienced.

Generalisability (*see also* external validity, applicability)

Refers to the extent to which a study's results provide a correct basis for generalisation beyond the setting of the study and the particular people studied. It implies the application of the results of a study to a group or population.

Gold standard

A method, procedure or measurement that is widely regarded or accepted as being the best available. Often used to compare with new methods.

Hazard ratio (HR)

When time to the outcome of interest is known, this is the ratio of the hazards in the treatment and control groups where the hazard is the probability of having the outcome at time *t*, given that the outcome has not occurred up to time *t*.

Heterogeneity

Refers to the differences in treatment effect between studies contributing to a meta-analysis. If there is significant heterogeneity, this suggests that the trials are not estimating a single common treatment effect.

Historical controls

Data from either a previously published series or previously treated patients at an institution that are used for comparison with a prospectively collected group of patients exposed to the technology or intervention of interest at the same institution.

Incidence

The number of new events (new cases of a disease) in a defined population, within a specified period of time.

Intention to treat (ITT)

An analysis of a clinical trial where participants are analysed according to the group to which they were initially randomly allocated, regardless of whether or not they dropped out, fully complied with the treatment, or crossed over and received the other treatment. By preserving the original groups one can be more confident that they are comparable.

Interaction

The relationship between a single variable (or covariate) and the treatment effect.

Intermediate outcomes

A true clinical endpoint that is not the ultimate endpoint of the disease but occurs quite late in the causal chain and represents manifestation of disease.

Interrupted time series

Treatment effect is assessed by comparing the pattern of (multiple) pretest scores and (multiple) post-test scores (after the introduction of the intervention) in a group of patients. This design can be strengthened by the addition of a control group which is observed at the same points in time but the intervention is not introduced to that group. This type of study can also use multiple time series with staggered introduction of the intervention.

Intervention

An intervention will generally be a therapeutic procedure such as treatment with a pharmaceutical agent, surgery, a dietary supplement, a dietary change or psychotherapy. Some other interventions are less obvious, such as early detection (screening), patient educational materials, or legislation. The key characteristic is that a person or their environment is manipulated in order to benefit that person.

Level of evidence

A hierarchy of study evidence that indicates the degree to which bias has been eliminated in the study design.

Meta-analysis (*see also* cumulative meta-anaysis)

Results from several studies, identified in a systematic review, are combined and summarised quantitatively.

Meta-regression

The fitting of a linear regression model with an estimate of the treatment effect as the dependent variable and study level descriptors as the independent variables.

Nonrandomised cross-over design

Participants in a trial are measured before and after introduction or withdrawal of the intervention and the order of introduction and withdrawal is not randomised.

Null hypothesis

The hypothesis that states that there is no difference between two or more interventions or two or more groups (eg males and females). The null hypothesis states that the results observed in a study (eg the apparent beneficial effects of the intervention) are no different from what might have occurred as a result of the operation of chance alone.

Number needed to harm (NNH) (*see also* number needed to treat)

When the treatment increases the risk of the outcome, then the inverse of the absolute risk reduction is called the number needed to harm.

Number needed to treat (NNT) (*see also* number needed to harm)

When the treatment reduces the risk of specified adverse outcomes of a condition, NNT is the number of patients with a particular condition who must receive a treatment for a prescribed period in order to prevent the occurrence of the adverse outcomes. This number is the inverse of the absolute risk reduction.

Observational studies

Also known as epidemiological studies. These are usually undertaken by investigators who are not involved in the clinical care of the patients being studied, and who are not using the technology under investigation in this group of patients.

Odds ratio (OR)

Ratio of the odds of the outcome in the treatment group to the corresponding odds in the control group.

Patient expected event rate (PEER)

The probability that a patient will experience a particular event (eg a stroke or myocardial infarction) if left untreated. Also known as baseline risk.

Patient-relevant outcome

Any health outcome that is meaningful to the patient. It can be the best surrogate outcome, resources provided as part of treatment, impact on productivity (indirect) or one that cannot be measured (eg pain, suffering). Common examples include: primary clinical outcomes, quality of life and economic outcomes.

Post-test only study

Patients undergo new technology and outcomes are described. This does not allow any comparisons.

Pretest/post-test study

Outcomes (pain, symptoms, etc) are measured in patients before receiving the new technology and the same outcomes are measured after. 'Improvement' in the outcome is reported. Often referred to as before-and-after studies.

Precision

A measure of how close the estimate is to the true value. It is defined as the inverse of the variance of a measurement or estimate. It is related to the *P*-value (the smaller the *P*-value, the greater the precision). (Also called statistical precision.)

Prevalence

Prevalence is a measure of the proportion of people in a population who have some attribute or disease at a given point in time or during some time period.

Prognostic model

A statistical model which estimates the patient's probability of developing the disease or outcome of interest from values of various characteristics (such as age, gender, risk factors).

Pseudorandomised controlled study

An experimental comparison study in which subjects are allocated to treatment/intervention or control/placebo groups in a nonrandom way (such as alternate allocation, allocation by day of week, odd–even study numbers, etc). These groups may therefore differ from each other in ways other than the presence of the intervention being tested. This contrasts to 'true' experiments (RCTs) where the outcomes are compared for groups formed by random assignment (and are therefore equivalent to each other in all respects except for the intervention).

Publication bias

Bias caused by the results of a trial being more likely to be published if a statistically significant benefit of treatment is found.

***P*-value** (*see also* confidence interval, precision, statistically significant effect)

The probability (obtained from a statistical test) that the null hypothesis (that there is no treatment effect) is incorrectly rejected.

NOTE: The *P*-value is often misunderstood. It does not, as commonly believed, represent the probability that the null hypothesis (that there is no treatment effect) is true (a small *P*-value therefore being desirable). The *P*-value obtained from a statistical test corresponds to the probability of claiming that there is a treatment effect when in fact there is no real effect.

Quality of evidence (*see also* strength of evidence)

Degree to which bias has been prevented through the design and conduct of research from which evidence is derived.

Quality of life

The degree to which persons perceive themselves able to function physically, emotionally and socially. In a more 'quantitative' sense, an estimate of remaining life free of impairment, disability, or handicap as captured by the concept of quality-adjusted life-years (QALYs).

Random error

The portion of variation in a measurement that has no apparent connection to any other measurement or variable, generally regarded as due to chance.

Randomisation

A process of allocating participants to treatment or control groups within a controlled trial by using a random mechanism, such as coin toss, random number table, or computer-generated random numbers. Study subjects have an equal chance of being allocated to an intervention or control group thus the two groups are comparable.

Randomised controlled trial

An experimental comparison study in which participants are allocated to treatment/intervention or control/placebo groups using a random mechanism, such as coin toss, random number table, or computer-generated random numbers. Participants have an equal chance of being allocated to an intervention or control group and therefore allocation bias is eliminated.

Randomised cross-over trial

Patients are measured before and after exposure to different technologies (or placebo) which are administered in a random order (and usually blinded).

Relative risk or risk ratio (RR)

Ratio of the proportions in the treatment and control groups with the outcome. This expresses the risk of the outcome in the treatment group relative to that in the control group.

Relative risk reduction (RRR)

The relative reduction in risk associated with an intervention. This measure is used when the outcome of interest is an adverse event and the intervention reduces the risk. It is calculated as one minus the relative risk, or:

$$\text{RRR} = 1 - (\text{event rate in treatment group} / \text{event rate in control group})$$

Relevance

The usefulness of the evidence in clinical practice, particularly the appropriateness of the outcome measures used.

Reliability

Also called consistency or reproducibility. The degree of stability that exists when a measurement is repeatedly made under different conditions or by different observers.

Risk difference (RD)

The difference (absolute) in the proportions with the outcome between the treatment and control groups. If the outcome represents an adverse event (such as death) and the risk difference is negative (below 0) this suggests that the treatment reduces the risk —referred to as the absolute risk reduction.

Selection bias

Error due to systematic differences in characteristics between those who are selected for study and those who are not. It invalidates conclusions and generalisations that might otherwise be drawn from such studies.

Size of effect

Refers to the size (or the distance from the null value indicating no treatment effect) of the summary measure (or point estimate) of the treatment effect and the inclusion of only clinically important effects in the 95% confidence interval.

Statistically significant effect (*see also* clinically important effect)

An outcome for which the difference between the intervention and control groups is statistically significant (ie the *P*-value is 0.05). A statistically significant effect is not necessarily clinically important.

Statistical precision (*see* precision)

Strength of evidence

The strength of evidence for an intervention effect includes the level (type of studies), quality (how well the studies were designed and performed to eliminate bias) and statistical precision (*P*-value and confidence interval).

Surrogate outcome

Physiological or biochemical markers that can be relatively quickly and easily measured and that are taken as predictive of important clinical outcomes. They are often used when observation of clinical outcomes requires longer follow-up. Also called intermediate outcome.

Systematic review

The process of systematically locating, appraising and synthesising evidence from scientific studies in order to obtain a reliable overview.

Time series

A set of measurements taken over time. An interrupted time series is generated when a set of measurements is taken before the introduction of an intervention (or some other change in the system), followed by another set of measurements taken over time after the change.

Type I error

When the null hypothesis (that there is no treatment effect) is incorrectly rejected.

Type II error

When the null hypothesis (that there is no treatment effect) is not rejected, but is actually false.

Validity

Of measurement: an expression of the degree to which a measurement measures what it purports to measure; it includes construct and content validity.

Of study: the degree to which the inferences drawn from the study are warranted when account is taken of the study methods, the representativeness of the study sample, and the nature of the population from which it is drawn (internal and external validity, applicability, generalisability).

Variance

A measure of the variation shown by a set of observations, defined by the sum of the squares of deviation from the mean, divided by the number of degrees of freedom in the set of observations.

ACRONYMS AND ABBREVIATIONS

ACE	angiotensin converting enzyme
ADAS	Alzheimer's Disease Assessment Scale
AIDS	acquired immune deficiency syndrome
CI	confidence interval
CIBIC	clinicians interview-based impression of change
DCCT	Diabetes Control and Complications Trial
ECG	Electrocardiogram
FOBT	faecal occult blood test
HAC	Health Advisory Committee, NHMRC
HIV	human immunodeficiency virus
HR	hazard ratio
ISIS	International Study of Infarct Survival
NHMRC	National Health and Medical Research Council
NIDDM	noninsulin dependent diabetes mellitus (or type II diabetes mellitus)
NNH	number needed to harm
NNT	number needed to treat
NSAID	nonsteroidal anti-inflammatory drug
OR	odds ratio
PEER	patient's expected event rate
QALY	quality-adjusted life-year
QoL	quality of life
RCT	randomised controlled trial
RD	risk difference
RR	relative risk/risk ratio

REFERENCES

ACCSG (American–Canadian Co-operative Study Group) (1983). Persantin aspirin trial in cerebral ischemia. Part II: Endpoint results. Stroke 16:406–415.

Antman EM, Lau J, Kupelnick B et al (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. Journal of the American Medical Association 268:240–248.

Barrett-Connor E (1991). Post-menopausal estrogen and prevention bias. Annals of Internal Medicine 115:455–456.

Berlin JA, Colditz GA (1990). A meta-analysis of physical activity in the prevention of coronary heart disease. American Journal of Epidemiology 132:612–628.

Berry G (1986). Statistical significance and confidence intervals. Medical Journal of Australia 144:618–619.

Boissel JP (1998). Individualizing aspirin therapy for prevention of cardiovascular events. Journal of the American Medical Association 280:1949–1950.

Boissel JP, Collet JP, Levre M et al (1993). An effect model for the assessment of drug benefit: example of antiarrthythmic drugs in postmyocardial infarction patients. Journal of Cardiovascular Pharmacology 22:356–363.

Bousser MG, Eschwege E, Haguenau M et al (1983). ‘AICLA’ controlled trial of aspirin and dipyridamole in the secondary prevention of athero-thrombotic cerebral ischaemia. Stroke 14:5–14.

Brand R, Kragt H (1992). Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. Statistics in Medicine 11:2077–2082.

Byers T, Perry G (1992). Dietary carotenes, vitamin C, and vitamin E as protective antioxidants in human cancers. Annual Reviews in Nutrition 12:139–159.

Cappelleri JC, Ioannidis JPA, Schmid CH et al (1996). Large trials vs meta-analysis of smaller trials. How do their results compare? Journal of the American Medical Association 276:1332–1338.

CAPRIE Steering Committee (1996). A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet 348:1329–1338.

CAPS (Cardiac Arrhythmia Pilot Study Investigators) (1988). Effects of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS. American Journal of Cardiology 61:501–509.

Carson CA, Fine MJ, Smith MA et al (1994). Quality of published reports of the prognosis of community acquired pneumonia. Journal of General Internal Medicine 9:13–19.

CAST (Cardiac Arrythmia Supression Trial) Investigators (1989). Preliminary report: effect of Encainide and Flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. New England Journal of Medicine 321:406–412.

Colditz GA, Miller JN, Mosteller F (1989). How study design affects outcomes in comparisons of therapy. I: Medical. Statistics in Medicine 8:441–454.

DCCT (Diabetes Control and Complications Trial) Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long term complications in insulin dependent diabetes mellitus. New England Journal of Medicine 329:977–986.

Diamond GA, Forrester JS (1983). Clinical trials and statistical verdicts: probable grounds for appeal. Annals of Internal Medicine 98:385–394.

Diener, HC, Cunha L, Forbes C et al (1996). European Stroke Prevention Study 2. Dipyridamole and acetylsalicylic acid in the secondary prevention of stroke. Journal of Neurological Sciences 143:1–13.

Eddy DM (1990a). Clinical decision making: from theory to practice. Anatomy of a decision. Journal of the American Medical Association 263:441–443.

Eddy DM (1990b). Clinical decision making: from theory to practice. Comparing benefits and harms: the balance sheet. Journal of the American Medical Association 263:2493–2505.

Egger M, Davey-Smith G (1995). Misleading meta-analysis. Lessons from ‘an effective, safe, simple’ intervention that wasn’t. British Medical Journal

Fleming TR, DeMets DL (1996). Surrogate end points in clinical trials: are we being misled? Annals of Internal Medicine 125:605–613.

Forgie MA, Wells PS, Laupacis A et al (1998). Preoperative autologous donation decreases allogeneic transfusion but increases exposure to all red blood cell transfusion. Archives of Internal Medicine 158:610–616.

Gelber RD, Goldhirsch A (1987). Interpretation of results from subset analyses within overviews of randomized clinical trials. Statistics in Medicine 6:371–378.

Glasziou PP, Irwig LM (1995). An evidence based approach to individualising treatment. British Medical Journal 311:1356–1359.

Glasziou P, Guyatt GH, Dans A et al (1998). Applying the results of trials and systematic reviews to individual patients (Editorial). ACP Journal Club Nov/Dec:A15–17.

Grady D, Rubin SM, Petitti DB et al (1992). Hormone therapy to prevent disease and prolong life in postmenopausal women. Annals of Internal Medicine 117:1016–1037.

Greenhalgh T (1997). Papers that summarise other papers (systematic reviews and meta-analyses). British Medical Journal 315:672–675.

Gueyffier F, Boutitie F, Boissel JP et al (1997). Effect of antihypertensive drug treatment on cardiovascular outcomes in women and men. Annals of Internal Medicine 126:761–767.

Guyatt GH, Sackett DL, Sinclair JC et al (1995). Users guides to the medical literature. IX. A method for grading health care recommendations. Journal of the American Medical Association 274:1800–1804.

Gyorkos TW, Tannenbaum TN, Abrahamowicz M et al (1994). An approach to the development of practice guidelines for community health interventions. Canadian Journal of Public Health (Supp 1);Jul–Aug:S9–13.

Hadorn DC, Baker D (1994). Development of the AHCPR-sponsored heart failure guideline: methodologic and procedural issues. Journal of Quality Improvement 20:539–554.

Hanley JA, Lippman-Hand A (1983). If nothing goes wrong is everything all right? Journal of the American Medical Association 249:1743–1744.

He J, Whelton PK, Vu B et al (1998). Aspirin and risk of hemorrhagic stroke. A meta-analysis of randomized controlled trials. Journal of the American Medical Association 280:1930–1935.

Hennekens CH, Buring JE, Manson JE (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. New England Journal of Medicine 334:1145–1149.

Henry D, Lim L, Rodriguez L et al (1996). Variability in risk of gastrointestinal complications with individual non-steroidal anti-inflammatory drugs: results of a collaborative meta-analysis. British Medical Journal 312:1563–1566.

Horwitz RI, Feinstein AR (1979). Methodologic standards and contradictory results in case-control research. American Journal of Medicine 66:556–564.

Hulley S, Grady D, Bush T et al (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Journal of the American Medical Association 280:605–613.

Hunt DL, McKibbon KA (1997). Locating and appraising systematic reviews. Annals of Internal Medicine 126:532–538.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 2:351–360.

ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. Lancet 345:669–685.

Jadad AR, Moore A, Carroll D et al (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clinical Trials 17:1–12.

Knapp MJ, Knopman DS, Solomon PR et al (1994). A 30-week randomized controlled trial of high-dose tacrine in patients with Alzheimer's disease. Journal of the American Medical Association 271:985–991.

Krum H, McNeil JJ (1998). The short life and rapid death of a novel antihypertensive and antianginal agent. What can we learn from the mibefradil experience? (Editorial). Medical Journal of Australia 169:408–409.

Lichtenstein MJ, Mulrow CD, Elwood PC (1987). Guidelines for reading case-control studies. Journal of Chronic Diseases 40:893–903.

Liddle J, Williamson M, Irwig L (1996). Improving Health Care and Outcomes. Method for Evaluating Research and Guideline Evidence. NSW Department of Health.

Lubsen J, Tijssen JG (1989). Large trials with simple protocols: indications and contraindications. Controlled Clinical Trials 10(Suppl):151S–160S.

Marshall KG (1996a). Prevention. How much harm? How much benefit? 1. Influence of reporting methods on perception of benefits. Canadian Medical Association Journal 154:1493–1499.

Marshall KG (1996b). Prevention. How much harm? How much benefit? 2. Ten potential pitfalls in determining the clinical significance of benefits. Journal of the Canadian Medical Association 154:1837–1843.

McGettigan P, Sly K, O'Connell D et al (1999). The effect of information framing on the practices of physicians. A systematic review of the published literature. Journal of General Internal Medicine 14:633–642.

McIntosh M (1996). The population risk as an explanatory variable in research synthesis of clinical trials. Statistics in Medicine 15:1713–1728.

Mellors JW, Munoz A, Giorgi JV et al (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Annals of Internal Medicine 126:946–954.

Moher D, Jadad AR, Tugwell P (1996). Assessing the quality of randomized controlled trials. Current issues and future directions. International Journal of Technology Assessment in Health Care 12:195–208.

Moher D, Jones A, Cook DJ et al (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609–613.

Mullins ME, Horowitz BZ, Linden DHJ et al (1998). Life-threatening interaction of mibefradil and beta blockers with dihydropyridine calcium channel blockers. Journal of the American Medical Association 280:157–158.

NEEGDP (North of England Evidence-based Guideline Development Project) (1997). Evidence based clinical practice guideline. ACE inhibitors in the primary care management of adults with symptomatic heart failure. Newcastle-upon-Tyne, UK: Centre for Health Services Research.

NHMRC (National Health and Medical Research Council) (1995). Guidelines for the Development and Implementation of Clinical Practice Guidelines. Canberra: AGPS.

NHMRC (1999). A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines. Canberra: NHMRC.

NHMRC (2000a). How to Review the Evidence: Systematic Identification and Review of the Scientific Literature. Canberra: NHMRC.

NHMRC (2000b). How to Compare the Costs and Benefits: Evaluation of the Economic Evidence. Canberra: NHMRC.

NZCSC (New Zealand Core Services Committee) (1995). Guidelines for the Management of Mildly Raised Blood Pressure in New Zealand. Wellington, New Zealand, Core Services Committee. (http://cebm.jr2.ox.ac.uk/docs/prognosis.html)

O'Brien WA, Hartigan PM, Daar ES et al (1997). Changes in plasma HIV RNA levels and CD4+ lymphocyte counts predict both response to antiretroviral therapy and therapeutic failure. Annals of Internal Medicine 126:939–945.

O'Connell D, Glasziou P, Hill S et al (1997). Assessing and increasing the generalisability of findings from randomised controlled trials (RCTs). Report to the United Kingdom National Health Service Health Technology Assessment Programme (unpublished).

O'Connell DL, Henry DA, Tomlins R (1999). Randomised controlled trial of effect of feedback on general practitioners' prescribing in Australia. British Medical Journal 318:507–511.

Omenn GS, Goodman GE, Thornquist MD et al (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. New England Journal of Medicine 334:1150–1155.

Powell KE, Thompson PD, Caspersen CJ et al (1987). Physical activity and the incidence of coronary heart disease. Annual Review of Public Health 8:253–287.

Realini JP, Goldzieher JW (1985). Oral contraceptives and cardiovascular disease: a critique of the epidemiologic studies. American Journal of Obstetrics and Gynecology 152:729–798.

Rogers SL, Doody RS, Mohs RC, Friedhoff LT (1998a). Donepezil improves cognition and global function in Alzheimer's disease: a 15-week, double-blind, placebo-controlled study. Donepezil Study Group. Archives of Internal Medicine 158:1021–1031.

Rogers SL, Farlow MR, Doody RS et al (1998b). A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Donepezil Study Group. Neurology 50:136–145.

Rothwell PM (1995). Can overall results of clinical trials be applied to all patients? Lancet 345(8965):1616–1619.

Sano M, Ernesto C, Thomas RG et al (1997). A controlled trial of selegiline, alpha-tocopherol, or both as treatment for Alzheimer's disease. New England Journal of Medicine 336:1216–1222.

Schmid CH, Lau J, McIntosh MW, Cappelleri JC (1998). An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Statistics in Medicine 17:1923–1942.

Schuchter L, Schultz DJ, Synnestvedt M et al (1996). A prognostic model for predicting 10-year survival in patients with primary melanoma. Annals of Internal Medicine 125:369-375.

Schulz KF, Chalmers I, Hayes RJ et al (1995). Empirical Evidence of Bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association 273:408–412.

Sharp S, Thompson SG, Douglas GA (1996). The relation between treatment benefit and underlying risk in meta-analysis. British Medical Journal 313: 735–738.

Simoons ML, Arnold AER (1993). Tailored thrombolytic therapy. A perspective. Circulation 88:2556–2564.

Simoons ML, Maggioni AP, Knatterud G et al (1993). Individual risk assessment for intracranial haemorrhage during thrombolytic therapy. Lancet 342: 1523–1528.

Sonis J, Doukas D, Klinkman M et al (1998). Applicability of clinical trial results to primary care. Journal of the American Medical Association 280:1746.

SPAF (Stroke Prevention in Atrial Fibrillation) Investigators (1994). Warfarin versus aspirin for prevention of thromboembolism in atrial fibrillation: stroke prevention in atrial fibrillation II study. Lancet 343:687–691.

TCCPSG (Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group) (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. New England Journal of Medicine 330:1029–1035.

Temple RJ (1995). A regulatory authority’s opinion about surrogate endpoints. In: Clinical Measurement in Drug Evaluation. Nimmo WS, Tucker GT (eds). New York: J Wiley.

Temple R (1999). Are surrogate markers adequate to assess cardiovascular disease drugs? Journal of the American Medical Association 282:790–795.

Thijs L, Fagard R, Lijnen P et al (1994). Why is antihypertensive drug therapy needed in elderly patients with systolodiastolic hypertension? Journal of Hypertension. Supplement 12(6):S25–34.

Thompson SG, Smith TC, Sharp SJ (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis. Statistics in Medicine 16:2741–2758.

Towler B, Irwig L, Glasziou P et al (1998). A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, Hemoccult. British Medical Journal 317:559–565.

Walter SD (1997). Variation in baseline risk as an explanation of heterogeneity in meta-analysis. Statistics in Medicine 16:2883–2900.

Wardlaw JM, Warlow CP, Counsell C (1997). Systematic review of evidence on thrombolytic therapy for acute ischaemic stroke. Lancet 350:607–614.

Whitehouse P (1998). Measurements of quality of life in dementia. In: Health Economics of Dementia. Wimo A, Jonsson B, Karlsson, Winblad B (eds). Chichester: John Wiley & Sons.

Yusuf S, Koon T, Woods K (1993). Intravenous magnesium in acute myocardial infarction. An effective, safe, simple and inexpensive intervention. Circulation 87:2043–2046.